8

Exploring spatial patterning in data values

8.1 Introduction

This chapter introduces a variety of methods for the analysis of spatial variation in single and multiple variables. Methods are introduced that allow for the exploration in changes in values from place to place or in the way in which variables are related. In the first case, an example problem might be to ascertain if zones tend to be more similar to their neighbours in some parts of the study area than in others (e.g. do neighbourhoods in some areas have similar characteristics while neighbourhoods in other areas are quite different). In the second case, we may want to address questions such as 'How does the relationship between altitude and precipitation vary from place to place?' (e.g. does altitude seem to have an effect on precipitation amount in some areas but not others). The initial focus of the chapter is on the analysis of spatial structure (spatial autocorrelation, i.e. the degree to which values at one location are similar to values at neighbouring locations). Next, the chapter moves on to computation of local statistics. The initial concern is with univariate measures; next, regression and correlation procedures are outlined that enable exploration of spatial relations between multiple variables. Finally, some other approaches are mentioned briefly before the chapter is summarized.

8.2 Spatial autocorrelation

Section 4.8 introduced the concepts of spatial autocorrelation and spatial dependence. Recall that spatial autocorrelation refers to the nature of correlation between neighbouring values while spatial dependence suggests the case where neighbouring values

are similar (positive spatial autocorrelation specifically). A key tool for the analysis of spatial autocorrelation, the Moran's *I* coefficient, was introduced in Section 4.8. A locally derived version of Moran's *I* is detailed in Section 8.4.1. There are various other spatial autocorrelation measures that are applied widely. These include Geary's *C*, amongst others (see Bailey and Gatrell, 1995). Measures like Moran's *I* and Geary's *C* are conventionally used to explore spatial autocorrelation with neighbours of observations—that is, they enable assessment of the degree to which values tend to be similar to neighbouring values. It is also possible to explore how spatial autocorrelation varies with the distance separating observations (e.g. we can use a geographical weighting approach). One very useful tool for the exploration of spatial autocorrelation is the Moran scatter plot. The Moran scatter plot relates individual values to weighted averages of neighbouring values and the slope of a regression line fitted to the points in the scatter plot gives global Moran's *I*. An application of the Moran scatter plot is detailed in the case study in Section 8.7.1.

8.3 Local statistics

Section 4.6 introduced the idea of moving windows, whereby any statistic could be computed locally using a subset of the data. Section 4.7 extended this idea through the concept of geographical weights and inverse distance weighted prediction was illustrated. Another geographically weighted approach was demonstrated in Section 7.3.2, which introduced kernel estimation. In the following section, the idea of locally derived statistics is explored further, with a particular focus on local measures of spatial autocorrelation. Section 8.5 outlines some approaches to exploring local variations in the relationships between different variables.

8.4 Local univariate measures

Standard univariate statistical measures are often computed within a moving window, as demonstrated previously in Section 4.6. Such measures allow exploration of the degree and nature of variation in summary statistics across the region of interest. For example, a local version of the standard deviation enables assessment of the degree of variability in the property of interest from place to place. Knowledge of such variation is often crucial in interpreting spatial data. Section 10.4 explores this issue further with a focus on raster grid data.

Geographical weighting schemes are widely used in the estimation of local statistics. The quartic kernel, illustrated in the previous chapter, includes bandwidth, τ, which determines the degree of weighting by distance. For a small bandwidth, locations close to the centre of the window will receive most of the weight. In contrast, for a large bandwidth, more distant locations will also receive quite large weights. A large

bandwidth therefore corresponds to a wide 'hump' and a small bandwidth corresponds to a narrow 'hump'. Another widely used weighting scheme is the Gaussian weighting scheme. As in previous chapters, the weight for location i can be given by w_{ij}, indicating the weight of sample i with respect to location j. The Gaussian weighting scheme is given by (Fotheringham *et al.*, 2002):

$$w_{ij} = \exp\left[-0.5\left(\frac{d_{ij}}{\tau}\right)^2\right] \tag{8.1}$$

which indicates that the weight for location i with respect to location j is obtained by multiplying –0.5 by the square of the distance d between locations i and j (i.e. d_{ij}) divided by the bandwidth τ and then obtaining the exponential value of the product (this can easily be computed in a spreadsheet package, where 'exp' is the standard abbreviation for the exponential function). As an example, for a bandwidth of 10 and a distance of 15:

$$\exp\left[-0.5\left[\frac{15}{10}\right]^2\right] = \exp[-0.5(1.5)^2] = \exp[-0.5 \times 2.25] = \exp[-1.125] = 0.3247$$

where exp[–1.125] can be obtained with $2.718281828^{-1.125}$ ($= 1/2.718281828^{1.125}$) and 2.718281828 is the approximate base of the natural logarithm (see Wilson and Kirkby (1980) for more details). Appendix B shows one way to compute the exponential function.

The Gaussian weighting scheme, for bandwidths of 5, 10, and 15 units, is illustrated in Figure 8.1. Note that, unlike the case for the quartic kernel, the bandwidth for the Gaussian weighting scheme does not extend to the outer edge of the kernel, but of course the bandwidth still determines the kernel size.

Any standard statistic can be geographically weighted (see Fotheringham *et al.* (2002) for more information). As an example of this geographical weighting scheme in practice, obtaining the locally weighted mean using the Gaussian weighting scheme is illustrated below. Following Section 4.7, the locally weighted mean is given by:

$$\overline{z}_i = \frac{\sum_{j=1}^{n} z_j w_{ij}}{\sum_{j=1}^{n} w_{ij}} \tag{8.2}$$

Table 8.1 and Figure 8.2 detail the locations of a set of observations which are the same as those used in Section 4.7. However, in this case, the first observation is treated as known. The weights, obtained using the Gaussian weighting scheme detailed above (with bandwidths of 5, 10, and 15 units), are given for each distance. The weights for each location are then multiplied by the value at that location. As an example, following Equation 8.1 for a bandwidth of 10 the products of the multiplications are summed, giving a value of 136.733. The weight values are also summed, giving a value of 6.643. The weighted mean is then obtained by $136.733/6.643 = 20.582$. Given the 20 observations, the unweighted mean is 19.050.

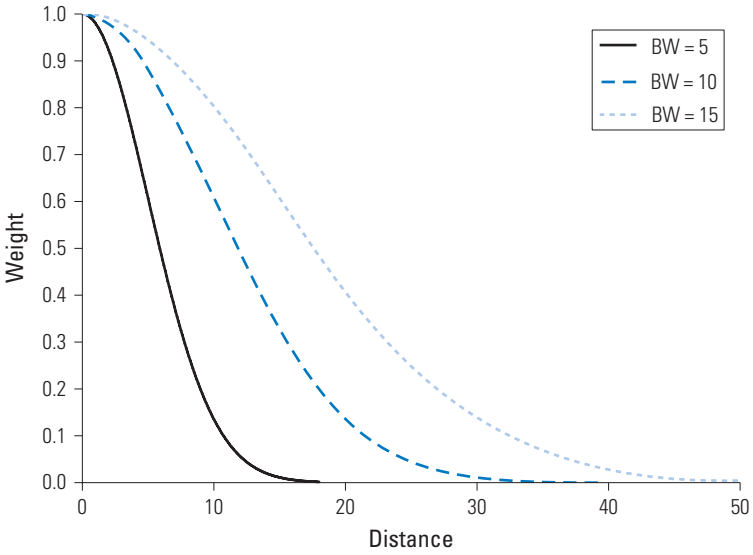

Figure 8.1 Gaussian weighting scheme: bandwidths of 5, 10, and 15 units.

Table 8.1 Observations (j), distance from observation 1 (d_{ij}), weights (w_{ij}) and weights multiplied by values ($z_j w_{ij}$)

			$\tau=5$		$\tau=10$		$\tau=15$	
j	d_{ij}	z_j	w_{ij}	$z_j w_{ij}$	w_{ij}	$z_j w_{ij}$	w_{ij}	$z_j w_{ij}$
1	0.000	9	1.000	9.000	1.000	9.000	1.000	9.000
2	4.404	14	0.679	9.499	0.908	12.706	0.958	13.409
3	9.699	43	0.152	6.553	0.625	26.867	0.811	34.889
4	10.408	12	0.115	1.375	0.582	6.981	0.786	9.433
5	10.871	34	0.094	3.198	0.554	18.829	0.769	26.147
6	12.958	26	0.035	0.905	0.432	11.230	0.689	17.903
7	13.959	24	0.020	0.487	0.377	9.059	0.649	15.565
8	14.066	33	0.019	0.631	0.372	12.271	0.644	21.260
9	15.506	34	0.008	0.277	0.301	10.219	0.586	19.927
10	17.256	10	0.003	0.026	0.226	2.256	0.516	5.160
11	17.606	8	0.002	0.016	0.212	1.698	0.502	4.017
12	18.018	13	0.002	0.020	0.197	2.564	0.486	6.319
13	18.025	11	0.002	0.017	0.197	2.167	0.486	5.344
14	18.285	24	0.001	0.030	0.188	4.510	0.476	11.416
15	19.253	9	0.001	0.005	0.157	1.410	0.439	3.949
16	21.335	15	0.000	0.002	0.103	1.541	0.364	5.455
17	23.845	14	0.000	0.000	0.058	0.816	0.283	3.957
18	23.988	34	0.000	0.000	0.056	1.914	0.278	9.465
19	24.464	3	0.000	0.000	0.050	0.150	0.264	0.793
20	24.522	11	0.000	0.000	0.049	0.544	0.263	2.891
	Sum	381	2.132	32.042	6.643	136.733	11.248	226.300
	Mean	19.050		15.032		20.582		20.119

Weights are obtained using the Gaussian weighting scheme with a bandwidth (τ) of 5, 10, and 15 units.

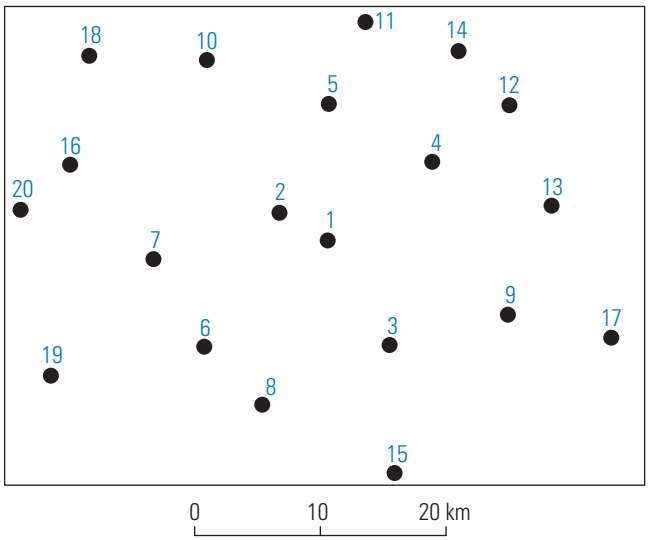

Figure 8.2 Locations of observations listed in Table 8.1.

With reference to Table 8.1, note how the weights for large distances are proportionately smaller when the bandwidth is smaller—that is, a small bandwidth gives most influence to close-by observations whereas with a large bandwidth more distant observations have proportionately larger weights. The weighted means (or other statistics) could be calculated anywhere—at the location of an observation or anywhere else (as for the inverse distance weighting example in Section 4.7). Fotheringham *et al.* (2002) discuss a range of geographically weighted statistics.

The spatial scale of a process can be explored using geographical weights. For example, by assessing the results obtained using a variety of different kernel bandwidths, it is possible to explore how much these results vary and, therefore, learn something about dominant scales of spatial variation. If the geographically weighted mean average changes a great deal as the bandwidth is increased for small bandwidths, but then stabilizes as the bandwidth is increased to some critical value, then we can say (with some caveats) that most variation in the property of concern is at some scale finer than that represented by the bandwidth distance at which results stabilize.

The following section shows how spatial autocorrelation measures can be computed locally.

8.4.1 Local spatial autocorrelation

In most published applications, spatial autocorrelation is measured over the entire study area, as was detailed in Section 4.8. However, such an approach masks any spatial variation in the spatial structure of the variable of interest. For this reason, various local measures of spatial autocorrelation have been developed. One of the most widely used is a local variant of Moran's *I* presented by Anselin (1995). It is given by:

$$I_i = z_i \sum_{j=1}^{n} w_{ij} z_j,\ j \neq i \tag{8.3}$$

where z_j are differences of variable y from its global mean $(y_i - \overline{y})$. In cases where zones are used (as opposed to points) the weights, w_{ij}, are often set to 1 for immediate neighbours of a zone and 0 for all other zones. Local I often appears in modified form:

$$I_i = \left[\frac{z_i}{s^2}\right] \sum_{j=1}^{n} w_{ij} z_j,\ j \neq i \tag{8.4}$$

where s^2 is the sample variance (the square of Equation 3.3). Note that local I values sum up to global Moran's I. Anselin (1995) describes an approach to testing for significant local autocorrelation based on random relocation of the data values, the objective being to assess if the observed configuration of values is significant. The GeoDa software offers the capacity to test the significance of local I using randomization[1] and to map significant clusters. Clusters are identified using the Moran scatter plot (Anselin, 1996).

Local I is demonstrated following Equation 8.4 using the following grid:

45 44 44
43 42 39
38 32 34

Local I is computed for the central cell and rook's case weighting is used—only the cells in the same row or column are used. The mean of values in the entire data set is needed first. Here there are only nine values and their mean is 40.111 and the sample variance is 21.861. Table 8.2 shows the original values (y_j), their deviations from the mean (z_j), the weights (w_{ij}), and weights multiplied by the deviations from the mean ($w_{ij}z_j$).

In this case the weights are row standardized, i.e. they sum to 1 (there are four values of 0.25 and these are for the four cells which share an edge with the central cell, which has a value of 42). z_i is 1.889, the sum of the weights multiplied by the deviations from the mean ($w_{ij}z_j$) is –0.611, as shown in Table 8.2. I_i is then given by $(^{1.889}/_{21.861}) \times -0.611 = -0.053$. In this case, I_i has a negative value, indicating negative spatial autocorrelation, i.e. neighbouring values tend to be dissimilar.

Figure 8.3 gives a map of the log of the number of persons per hectare in Northern Ireland in 2001; the values were logged as the raw population densities had a positively skewed distribution and the transformed values have almost zero skew. Figure 8.4 gives an example of the application of I_i for measuring spatial autocorrelation (using queen's case contiguity; the use of contiguity schemes with non-gridded data was outlined in Section 4.8) in logged population density given the data shown in

1 see https://www.geoda.uiuc.edu/support/help/glossary.html

Table 8.2 Values, differences from the mean, and weights

y_i	z_i	w_{ij}	$w_{ij}z_j$
45	4.889	0.000	0.000
43	2.889	0.250	0.722
38	–2.111	0.000	0.000
44	3.889	0.250	0.972
42	1.889	0.000	0.000
32	–8.111	0.250	–2.028
44	3.889	0.000	0.000
39	–1.111	0.250	–0.278
34	–6.111	0.000	0.000
Sum		1.000	–0.611

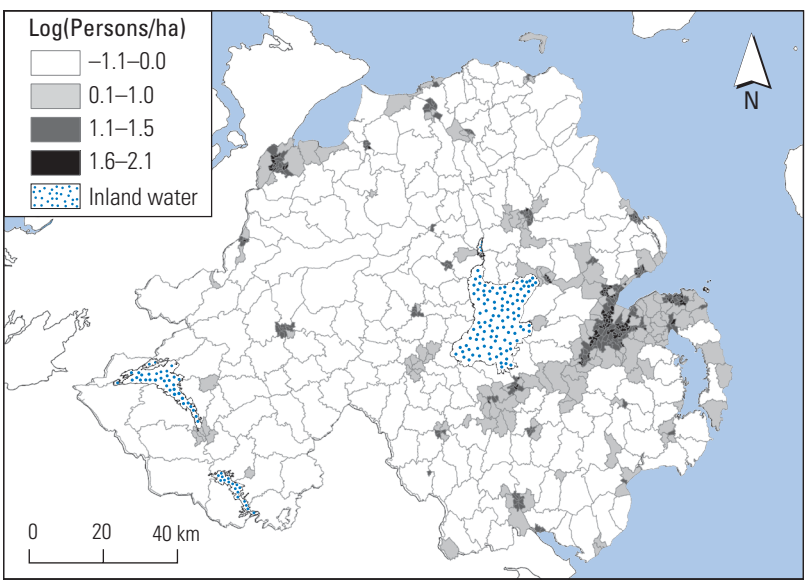

Figure 8.3 Log of persons per hectare in Northern Ireland in 2001. Northern Ireland Census of Population data—© Crown Copyright. Reproduced under the terms of the Click-Use Licence.

Figure 8.3. Global I was 0.665. In the case of I_i, there are large positive values in rural areas (which have larger zones, since zone size is a function of the population density). In urban areas like Belfast (in the east), the contrast between central areas with high population densities and suburban and rural areas with lower population densities are apparent. Areas with contrasting values are marked by small positive or negative values of I_i. Note that the results are a function of the particular form of zones used (see Section 4.9 for a relevant discussion). There are many published applications of local I (e.g. Anselin, 1995; Lloyd, 2006).

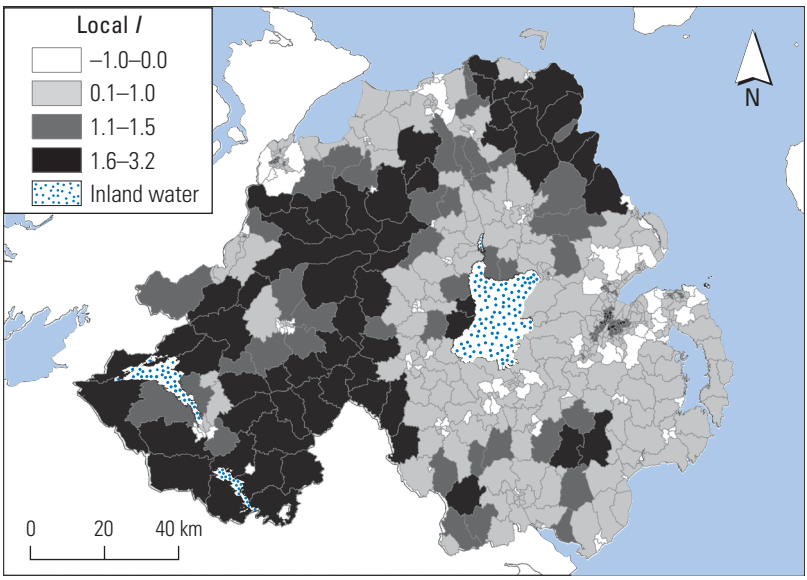

Figure 8.4 Local *I* for log of persons per hectare in Northern Ireland in 2001 using queen's case contiguity. Northern Ireland Census of Population data—© Crown Copyright. Reproduced under the terms of the Click-Use Licence.

The focus of the chapter now moves onto analysis of spatial patterning in the relationships between multiple variables.

8.5 Regression and correlation

The subject of regression and correlation was introduced in Section 3.3. Some regression-based analyses of spatially referenced variables map the residuals (in the two-variable case, this means the difference between the value indicated by the line of best fit and the observed value) from the regression (see the example in Figure 3.5). It is straightforward to take this a step further and explore not just how accurate fitted values are from place to place but to consider how the relationships between variables differ spatially. As with univariate statistics, multivariate approaches (such as correlation and regression) can be conducted in a moving window. The next section introduces some approaches to regression in the analysis of spatial data. Next, the focus is on regression conducted in a moving window. Following this, a geographically weighted approach is detailed and this account makes use of matrix algebra to obtain regression coefficients.

8.5.1 Spatial regression

An assumption of standard ordinary least squares regression is independence of observations. As discussed in Section 3.5, this assumption rarely holds true for spatial

data. Several approaches exist which take into account the spatial structure in variables and therefore allow for spatial dependence (e.g. Rogerson, 2006; Ward and Gleditsch, 2008). With generalized least squares (GLS) regression, information on spatial dependence can be utilized when estimating the regression coefficients. More specifically, the GLS regression coefficients can be estimated given information on the degree of similarity between variables as a function of the distance by which they are separated. Bailey and Gatrell (1995) provide an account of GLS. Spatial autoregressive models provide another means of accounting for spatial structure. Lloyd (2006) outlines the simultaneous autoregressive model, which includes an interaction parameter representing interactions between neighbouring observations. If the interaction parameter is unknown (the usual case), then such models cannot be fitted using ordinary least squares and specialist software is required. The simultaneous estimation of the interaction parameter and the regression (β) coefficients can be conducted using a maximum likelihood procedure, as described by Schabenberger and Gotway (2005). The GeoDa software of Anselin *et al.* (2006) allows for spatial autoregressive modelling. Such models help to overcome the problem of analysing relations between spatially referenced variables, but they provide only a single set of coefficients. Increasingly, in GIS contexts, studies take into account the local context. Local approaches entail estimating regression coefficients using either local data subsets or a geographical weighting scheme. Two local regression approaches are outlined next.

8.5.2 Moving window regression

Section 8.4.1 introduced the idea that spatial autocorrelation of variables is often observed to vary spatially. Local measures which take these variations into account may therefore be worthwhile. Similarly, relationships between variables may differ markedly across an area. As an example, many studies have shown that altitude and precipitation amount are related in some regions. However, while the two variables may be strongly related in some areas, a global regression of altitude and precipitation amount may demonstrate only a weak relationship (see Lloyd (2005) for a relevant case study). Some kind of local regression procedure is therefore needed to enable exploration of some geographically variable relationships.

One straightforward approach to exploring how relationships vary spatially is simply to conduct a standard regression in a moving window. In other words, regression is carried out using only the data in the moving window and the end result is a set of maps of regression coefficients. Moving window regression (MWR) has been used in various studies (see Lloyd (2005, 2006) and Lloyd and Shuttleworth (2005) for examples). MWR is identical to the regression procedure detailed in Section 3.3, the only difference being that regression is conducted for data subsets in a moving window rather than for all data simultaneously. Building on the geographical weighting principles outlined previously (see Section 4.7), this approach can be extended such that the influence of observations in the regression is decreased as distance from the centre of the moving window increases. Such an approach is the subject of the following section.

8.5.3 Geographically weighted regression

This book has outlined various geographically weighted statistics and argued that such approaches are intuitively sensible since we expect places close together to be more alike than places a greater distance apart. Geographically weighted regression (GWR) extends the same principle to regression analysis. GWR has become a core tool in many analyses of spatial data. The approach is described in detail by Fotheringham *et al.* (2002) but an account of some key principles is given here. Essentially, the key steps in a GWR analysis are as follows:

1. Go to a location (zone or point).
2. Conduct regression using all data (or some subset) but give greater weight (influence) to locations that are close to the location of interest—a geographical weighting scheme is used.
3. Move to the next location and go back to stage 2 until all locations have been visited.

The output is a set of regression coefficients (e.g. for bivariate regression (with one independent and one dependent variable), the intercept, and slope) at each location. GWR coefficients are obtained using:

$$\beta(\mathbf{x}_i) = (\mathbf{Y}^T\mathbf{W}(\mathbf{x}_i)\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{W}(\mathbf{x}_i)\mathbf{z} \tag{8.5}$$

This is the same as for ordinary unweighted regression (Equation 3.9), except that the regression coefficients are computed for each location $\mathbf{x}_i$ and there are geographical weights given by $\mathbf{W}(\mathbf{x}_i)$. If all weights were equal to 1 then this would correspond to standard unweighted regression. The weights matrix is given by:

$$\mathbf{W}(\mathbf{x}_i) = \begin{bmatrix} w_{i1} & 0 & 0 & 0 \\ 0 & w_{i2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & w_{in} \end{bmatrix}$$

w_{i1} is the weight given the distance between the location i and observation 1. The diagonal dots ($\ddots$) indicate that the matrix can be expanded—that is, if n (the number of observations) is 5 then the matrix will have 5×5 entries, with non-zero values only in the diagonal of the matrix.

One weighting scheme used widely in GWR contexts is the Gaussian weighting scheme, detailed in Equation 8.1. Note that MWR is a special case of GWR where the weights for the n nearest neighbours are set to 1 and all other weights are 0.

GWR is illustrated using the data listed in Table 8.3 and mapped in Figure 8.5. Table 8.3 gives the coordinates of the observations, variable 1 (independent) and variable 2 (dependent) values, distance from the first observation, and geographical weights ('Geog. wt.'; using the Gaussian weighting function) for a bandwidth of 10 units. The computation

Table 8.3 Coordinates of observations, variable 1 and 2 values, distance from the first observation, and geographical weights

No.	*x* coordinate	*y* coordinate	Variable 1 (*y*)	Variable 2 (*z*)	Distance (d_{ij})	Geog. wt. (w_{ij})
1	25.00	45.00	12	6	0.00	1.0000
2	25.51	44.14	34	52	1.00	0.9950
3	21.87	48.90	32	41	5.00	0.8825
4	27.60	52.57	12	25	8.00	0.7261
5	16.69	31.33	11	22	16.00	0.2780
6	42.52	35.35	14	9	20.00	0.0889
7	9.20	65.65	56	43	26.00	0.0340
8	29.23	76.72	75	67	32.00	0.0060
9	61.37	66.01	43	32	42.00	0.0002

Bandwidth = 10 units.

of geographical weights using Equation 8.1 was illustrated in Section 8.4. Note that the number of observations is small and this example is used purely for ease of illustration.

In this example, the weight matrix is therefore:

$$\mathbf{W}(\mathbf{x}_i) = \begin{bmatrix} 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.9950 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8825 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7261 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2780 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0889 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0340 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0060 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0002 \end{bmatrix}$$

The regression coefficients are obtained as for the standard regression approach detailed above except that $\mathbf{Y}^T$ is multiplied by $\mathbf{W}(x_i)$. Following this procedure (by referring to the example for global regression in Section 3.3 and Appendix E it should be possible to work out what is going on):

$$\mathbf{Y}^T\mathbf{W}(\mathbf{x}_i) = \begin{bmatrix} 1 & 0.9950 & 0.8825 & 0.7261 & 0.2780 & 0.0889 & 0.0340 & 0.0060 & 0.0001 \\ 12 & 33.8300 & 28.2400 & 8.7132 & 3.0580 & 1.2446 & 1.9040 & 0.4500 & 0.0086 \end{bmatrix}$$

$$\mathbf{Y}^T\mathbf{W}(\mathbf{x}_i)\mathbf{Y} = \begin{bmatrix} 4.0107 & 89.4484 \\ 89.4484 & 2494.2646 \end{bmatrix}$$

$$(\mathbf{Y}^T\mathbf{W}(\mathbf{x}_i)\mathbf{Y})^{-1} = \begin{bmatrix} 1.2454 & -0.0447 \\ -0.0447 & 0.0020 \end{bmatrix}$$

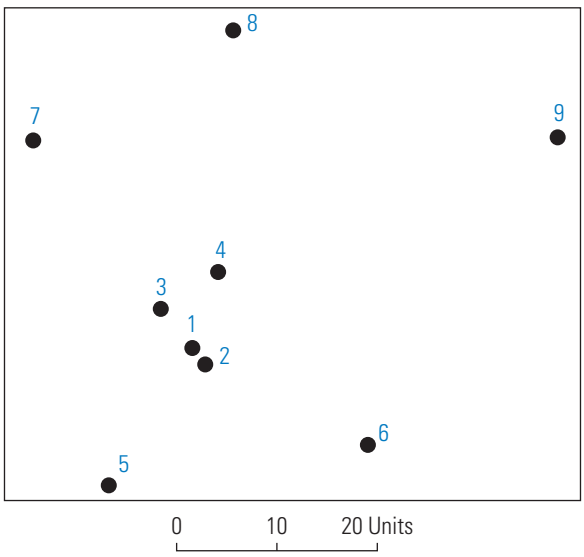

Figure 8.5 Locations of the data in Table 8.3.

$$\mathbf{Y}^T\mathbf{W}(\mathbf{x}_i)\mathbf{z} = \begin{bmatrix} 120.8615 \\ 3397.6046 \end{bmatrix}$$

$$\beta(\mathbf{x}_i) = \mathbf{Y}^T\mathbf{W}(\mathbf{x}_i)\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{W}(\mathbf{x}_i)\mathbf{z} = \begin{bmatrix} -1.223 \\ 1.406 \end{bmatrix}$$

The intercept, $\beta_0(\mathbf{x}_i)$, is therefore −1.223 and the slope, $\beta_1(\mathbf{x}_i)$, is 1.406. As an example, for the first case in Table 8.3, the fitted value is $-1.223+1.406\times 12=15.649$.

Figure 8.6 shows the ordinary (unweighted) and geographically weighted regression lines fitted with the geographical weights indicated. Note the effect of geographical weighting on, in effect, pulling the line towards the points with larger weights. You can match the geographical weight and the values of the two variables to the entries in Table 8.3.

The goodness of fit of the GWR locally can be assessed using the geographically weighted coefficient of determination (r^2) (Fotheringham *et al.*, 2002). The geographically weighted r^2 for location i is given by:

$$r_i^2 = \frac{\text{TSS}_i - \text{RSS}_i}{\text{TSS}_i} \tag{8.6}$$

where TSS_i is the geographically weighted total sum of squares given by:

$$\text{TSS}_i = \sum_{j=1}^{n} w_{ij}(z_j - \overline{z})^2$$

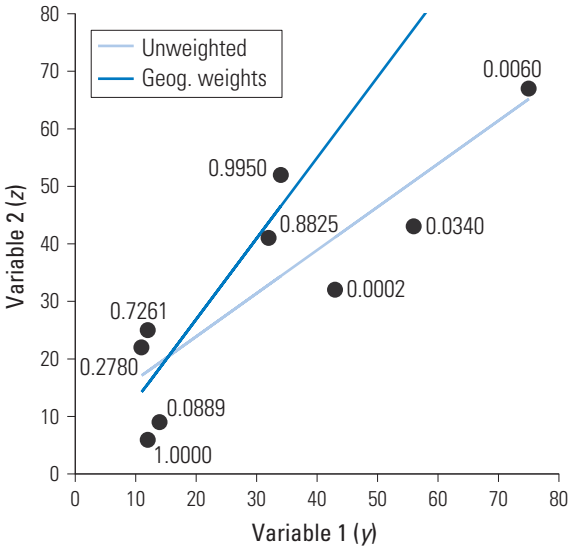

Figure 8.6 Regression using the data in Table 8.3: unweighted and geographically (geog.) weighted, with geographical weights indicated.

This is the sum of the weights multiplied by the squared difference between each (dependent variable) value and its mean. RSS_i is the geographically weighted residual sum of squares given by:

$$RSS_i = \sum_{j=1}^{n} w_{ij}(z_j - \hat{z}_j)^2$$

This is the sum of the weights multiplied by the squared residual (the difference between each value and the value given the GWR model).

The calculations following Equation 8.6 are presented in Table 8.4.

In this case, each term is as follows:

$$TSS_i = \sum_{j=1}^{n} w_{ij}(z_j - \overline{z})^2 = 1286.326$$

$$RSS_i = \sum_{j=1}^{n} w_{ij}(z_j - \hat{z}_j)^2 = 266.227$$

$$r_i^2 = \frac{TSS_i - RSS_i}{TSS_i} = \frac{1286.326 - 266.227}{1286.326} = 0.7930$$

The unweighted r^2 is 0.7413 and the geographically weighted r^2 is 0.7930 (or 0.7966 calculated using purpose-written software, the difference being due to rounding errors). In other words, the GWR model is a better fit than the unweighted model in this case and this suggests that taking into account distance from the location of

Table 8.4 Calculations for the geographically weighted coefficient of determination (data as for Table 8.3), with mean ($\overline{z}$) of 33

Obs. (j)	$w_{ij}(z_j - \overline{z})^2$	$\hat{z}_j$	$w_{ij}(z_j - \hat{z}_j)^2$
1	729.000	15.649	93.103
2	359.195	46.581	29.219
3	56.480	43.769	6.766
4	46.470	15.649	63.491
5	33.638	14.243	16.728
6	51.206	18.461	7.957
7	3.400	77.513	40.499
8	6.936	104.227	8.315
9	0.000	59.235	0.148
Sum	1286.326		266.227

Obs., observation.

interest is beneficial in this case. Again, it should be stressed that the number of observations is small for this example and this should be considered in any analysis.

The GWR software offers two approaches to assessing the significance of the GWR model. In essence, these approaches allow users to determine if any of the local parameter estimates are 'significantly non-stationary' (Fotheringham *et al.*, 2002, p. 213), where a non-stationary model is one which has parameters that vary geographically. In other words, the test enables assessment of the degree to which the GWR model is an improvement over a standard global model.

To illustrate GWR, data on elevation and precipitation amount in Northern Ireland for July 2006 (with data at 149 locations) are analysed. The data locations are shown in Figure 8.7 and an interpolated (see Section 9.2) map of precipitation amounts is given in Figure 9.7. Figure 8.8 is a scatter plot for all observations (it is a global regression) with a fitted regression line. This indicates that the expected precipitation amount at a location with an altitude of 0 m is 52.203 mm (i.e. the intercept) and that this increases on average by 0.1162 mm (the slope) with an increase in altitude of 1 m. While the coefficient of determination (r^2) indicates that the model explains some 42% of the variation, there is clearly much variation around the regression line and GWR allows for exploration of local variations in this relationship.

A GWR bandwidth can be selected using cross-validation. Cross-validation entails removal of an observation, regression conducted using the remaining observations, and prediction of the removed value—that is, the regression equation is used to predict the value of z (the dependent) given a value of y (the independent), as described in Section 3.3. The removed value is then added back and the observation removed at the next location (in whatever order the locations are visited) after which the procedure is repeated for all remaining observations. The difference between the observed and predicted values is then computed. The bandwidth that results in the smallest cross-validation error is retained. An additional method for bandwidth selection, which is employed by the GWR software of Fotheringham *et al.* (2002), is called the

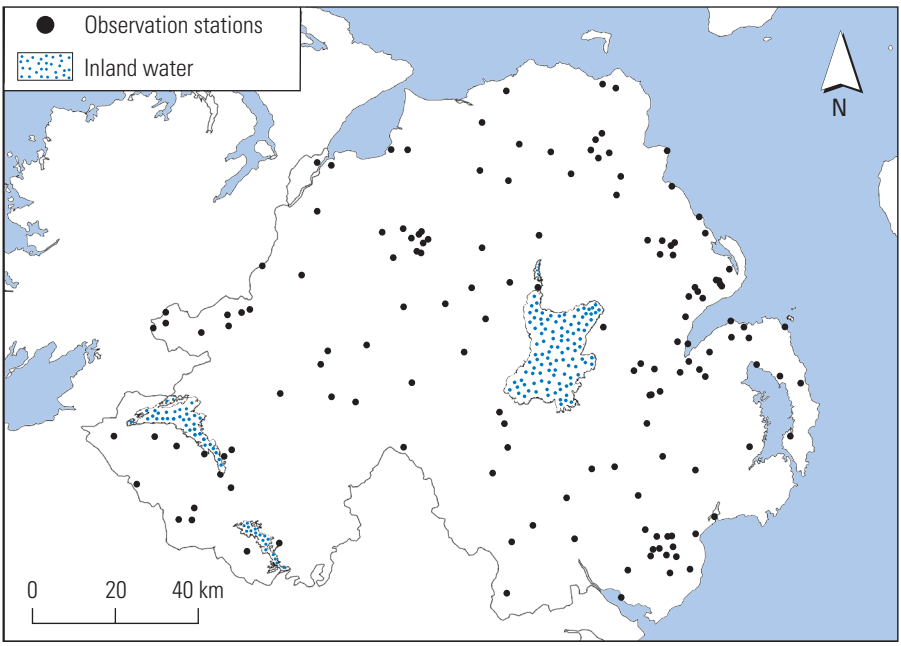

Figure 8.7 Locations of precipitation observations in July 2006.

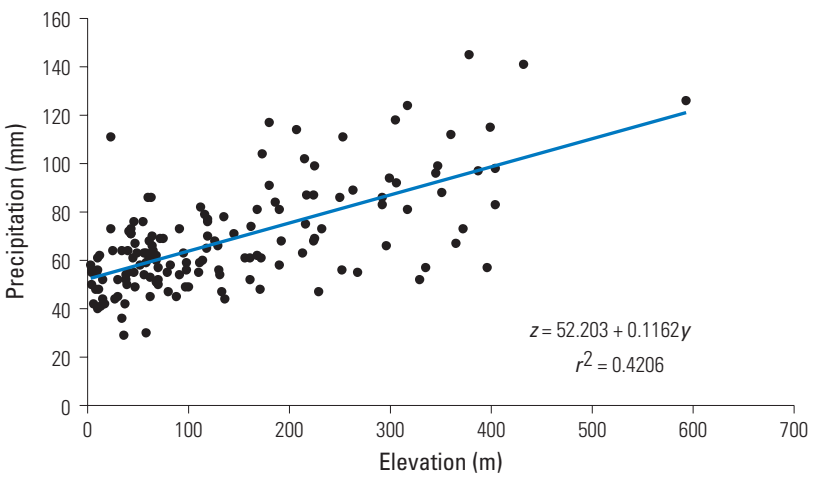

Figure 8.8 Elevation against July 2006 precipitation amount in Northern Ireland.

Akaike Information Criterion (AICc; Fotheringham *et al.*, 2002, Equation 4.21). The AICc allows for assessment of the number of degrees of freedom and the goodness of fit of the model and it can be used to compare the performance of a global regression model and GWR (Fotheringham *et al.*, 2002). The geographically weighted bandwidth of 9952.763 m was selected using the AICc in version 3.0 of the GWR software.

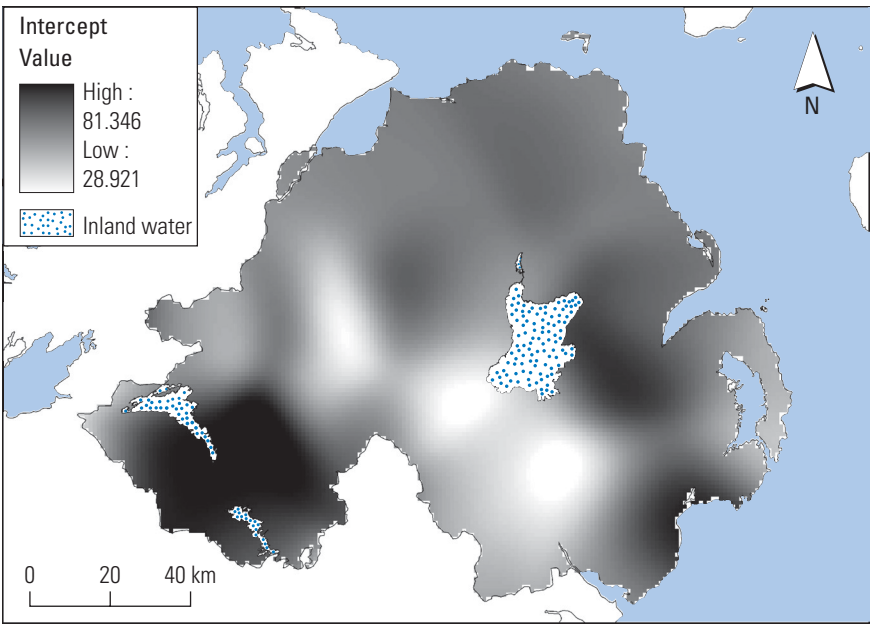

Figure 8.9 GWR intercept: elevation against July 2006 precipitation amount in Northern Ireland.

GWR can be conducted at any location, thus a GWR model can be fitted at a location where there is no observation. In the following example, GWR model parameters were obtained on a 1-km grid and this makes visualization of spatial variation in parameters easier than when parameter estimates are made at the locations of observations only. Figure 8.9 shows the mapped GWR intercept values, Figure 8.10 shows the slope coefficients, and Figure 8.11 shows the coefficients of determination (r^2; here computed outside of GWR 3.0, as that package does not return standard GW r^2 values). Note that the number of decimal places is kept at three for all of the following figures since the values for the slope and r^2 are small.

Clearly, there is much variation in the nature and the strength of the relationship between elevation and precipitation amount in Northern Ireland. Values of the GWR intercept (Figure 8.9) are largest (and larger than the global model intercept of 52.203 mm) in parts of the south west, parts of the mid east (south and west of Belfast), and the south east. Where the intercept is large and the slope has a negative or small positive value, this indicates that, in such areas, precipitation amounts are large (in proportion to the size of the intercept), irrespective of the elevation. Comparison of some areas in Figure 8.9 (GWR intercept) with Figure 8.10 (GWR slope) shows that some areas (e.g. in parts of the south west) fulfil these criteria. The GWR slope values (Figure 8.10) are largest in parts of the far west, the Ards Peninsula (south and west of Belfast), and some other areas, most notably to the south and west of Lough Neagh. Where the GWR r^2 (Figure 8.11) is also large, this suggests that elevations and precipitation amounts are strongly related (i.e. a large slope indicates a large increase in

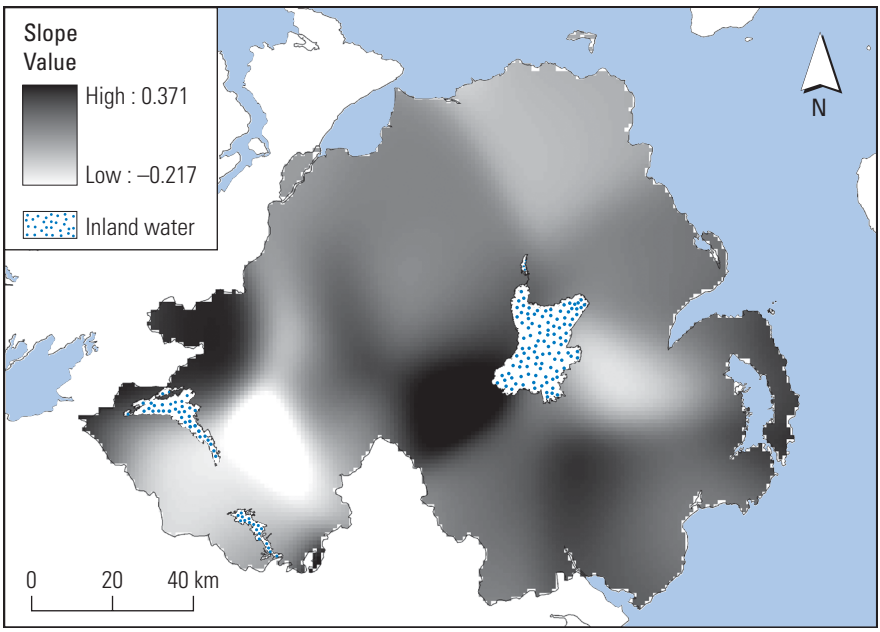

Figure 8.10 GWR slope parameter: elevation against July 2006 precipitation amount in Northern Ireland.

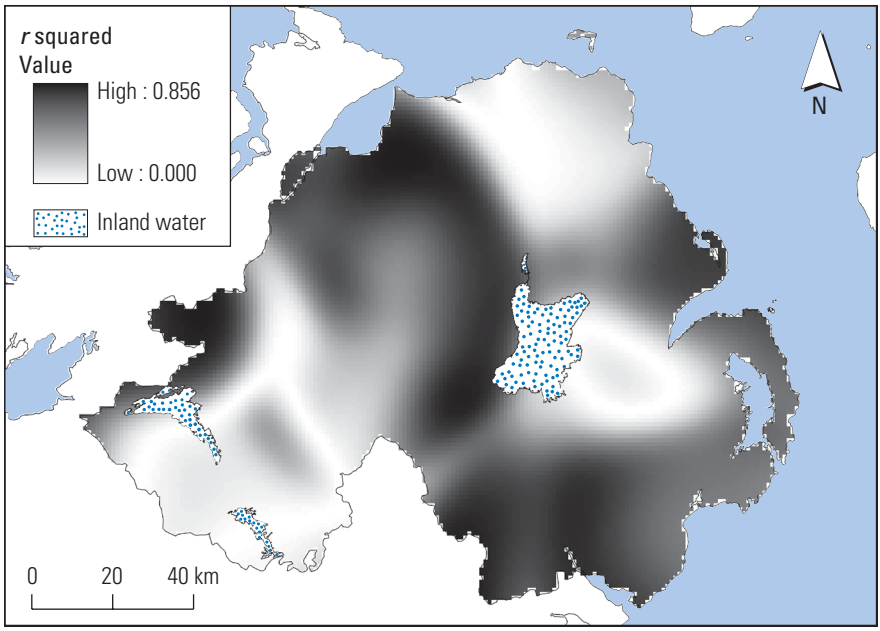

Figure 8.11 GWR coefficient of determination: elevation against July 2006 precipitation amount in Northern Ireland.

precipitation amount with an increase in elevation). Parts of the far west of the region have large slope and large r^2 values. The GWR slopes and r^2 values tend to be larger in areas with larger elevation values (see Figure 10.5 for a map of elevation in Northern Ireland) and this suggests that the elevation–precipitation relationship tends to be less strong (and therefore elevation is a less useful predictor) in areas with small elevation values. Trends in precipitation amounts are directional and are a function of many factors not considered here but, as the example demonstrates, GWR is a powerful means of exploring spatially variable relationships such as those between elevation and precipitation amount.

Relationships between many variables of interest in the physical and social sciences are a function of geography. These include the previous example of altitude and precipitation as well as other variables such as employment status and religion. Where such geographical variations are suspected, standard global regression analyses may be inadequate and an analysis based on the application of GWR may reveal a far richer picture than would be obtained through conventional regression analysis. GWR allows assessment of how far the nature of relationships (e.g. are variables related positively or negatively) vary spatially and how strongly variables are related in different regions. GWR has been used in many other contexts. Brunsdon *et al.* (2001) used GWR to explore the average elevation–precipitation relationship across Britain, while Lloyd and Shuttleworth (2005) used GWR to explore spatial variation in the relationship between commuting distance and a range of other variables in Northern Ireland. Some authors have commented on potential problems associated with GWR, particularly in terms of multicollinearity (i.e. strong correlations between independent variables). This may have an impact on the interpretation of the local regression coefficients (see Wheeler and Tiefelsdorf, 2005). In short, where there are multiple independent variables in the GWR analysis, and these independent variables are strongly related, the values and signs of coefficients for individual variables may be highly misleading. Methods for diagnosing collinearity and a solution to this problem are detailed by Wheeler (2007). Another issue which should be considered is the availability of enough observations with significantly non-negative weights for the purposes of GWR parameter estimation. One way around this problem is to use an adaptive bandwidth whereby the size of the bandwidth varies as a function of the density of observations in a given area (see Fotheringham *et al.*, 2002).

8.6 Other approaches

This chapter offers only a brief summary of some widely used approaches for the analysis of spatial structure in single and multiple variables. The selection is biased and many other approaches could have been included. For example, in terms of regression approaches, a body of models called multilevel models is used widely by geographers (and others) to explore relationships between variables at different spatial scales (see Fotheringham *et al.* (2002) and Lloyd (2006) for summaries of some other approaches).

The chapter introduces some key methods and concepts which will hopefully aid understanding of other approaches not considered here.

8.7 Case studies

This section comprises two case studies that make use of the same data set, which is provided on the book website. The data represent elevations in Switzerland and precipitation amounts for 8 May 1986; the number of observations is 467. The data are described by Dubois (2003) and the precipitation measurements are shown in Figure 8.12. The first case study makes use of Moran's *I* autocorrelation coefficient to explore spatial variation in precipitation amount. The second study uses GWR to explore spatial variation in the relationship between elevation and precipitation amount.

8.7.1 Spatial autocorrelation analysis

Moran's *I* was computed using geographical weights with the Gaussian weighting function defined in Equation 8.1. For a 10-km bandwidth, *I* was 0.722, indicating strong positive spatial autocorrelation in precipitation amounts. The data were further

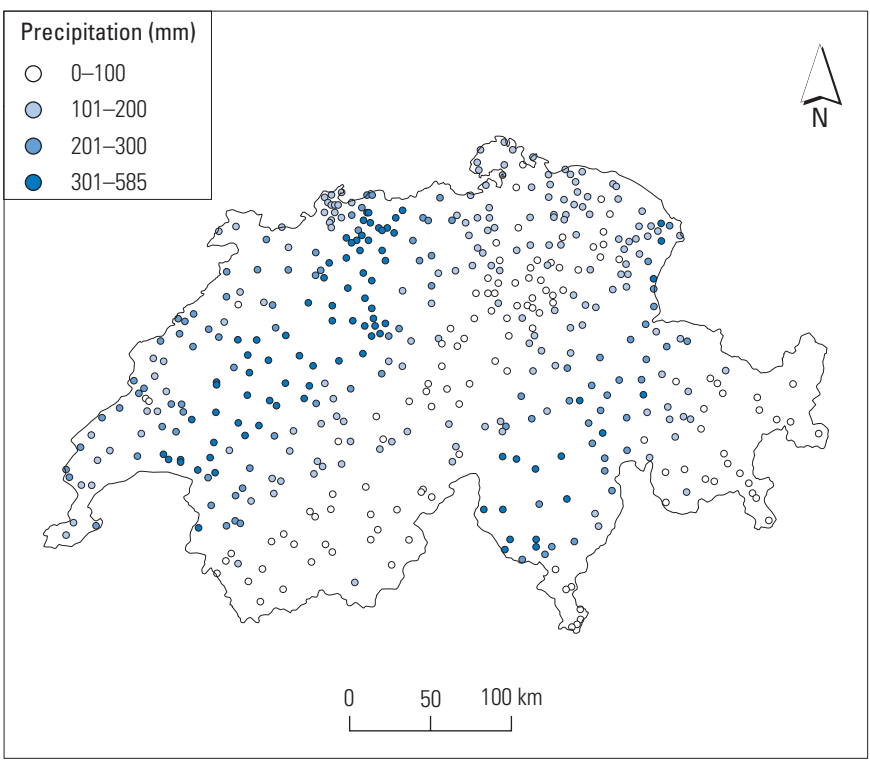

Figure 8.12 Precipitation measurements for 8 May 1986 in Switzerland.

interrogated through computing the Moran scatter plot, as described in Section 8.2, and this is shown in Figure 8.13. The plot shows the value at location *i* plotted against the weighted average of neighbouring values and the slope of the line fitted to the scatter plot gives global Moran's *I*. The plot allows assessment of outliers—those values that do not fit the general trend. For example, there are some points in the right-hand side of the upper right quadrant that fall far from the fitted line. Identifying these locations on a map could be informative, and this would be a sensible part of a fuller analysis. Moran's *I* could be computed using a variety of different bandwidths and changes in results with change in bandwidth help to suggest the scales over which the property is spatially autocorrelated.

An additional step that could be taken is to map the local *I* values, as in the case of Figure 8.4. Moran scatter plots and local *I* can both be computed using the GeoDa software (Anselin *et al.*, 2006).

8.7.2 Geographically weighted regression

For this analysis, the elevation data were provided as a digital elevation model (DEM) and the values at each precipitation observation location were extracted; the DEM is shown in Figure 10.17. A global regression of elevation and daily precipitation amount suggests that the two are only weakly related—the r^2 value was 0.0366 and the slope parameter coefficient was negative. Intuitively, we would expect elevation and precipitation to be positively related in at least some areas even for a period as short as a day (note that, for the example for Northern Ireland given in Section 8.5.3, the data were for monthly precipitation amounts). The GWR software (Fotheringham *et al.*, 2002)

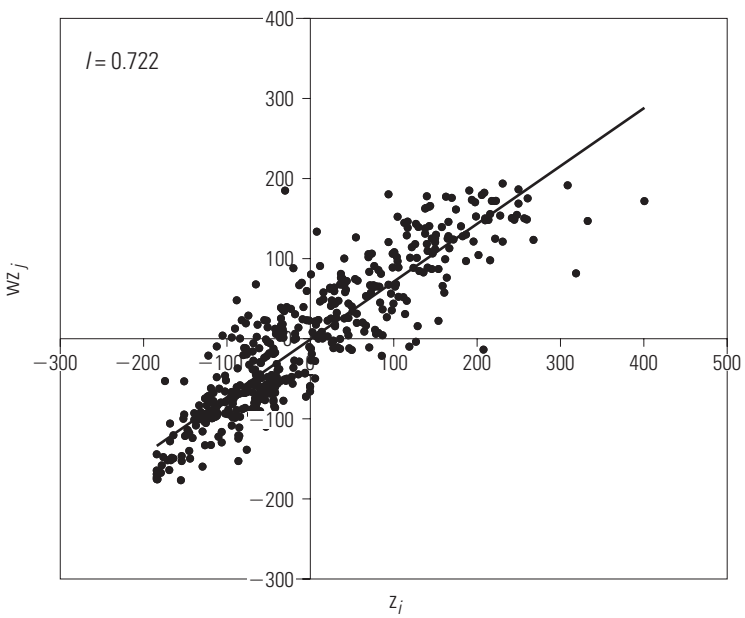

Figure 8.13 Moran scatter plot for precipitation measurements: 10-km Gaussian bandwidth.

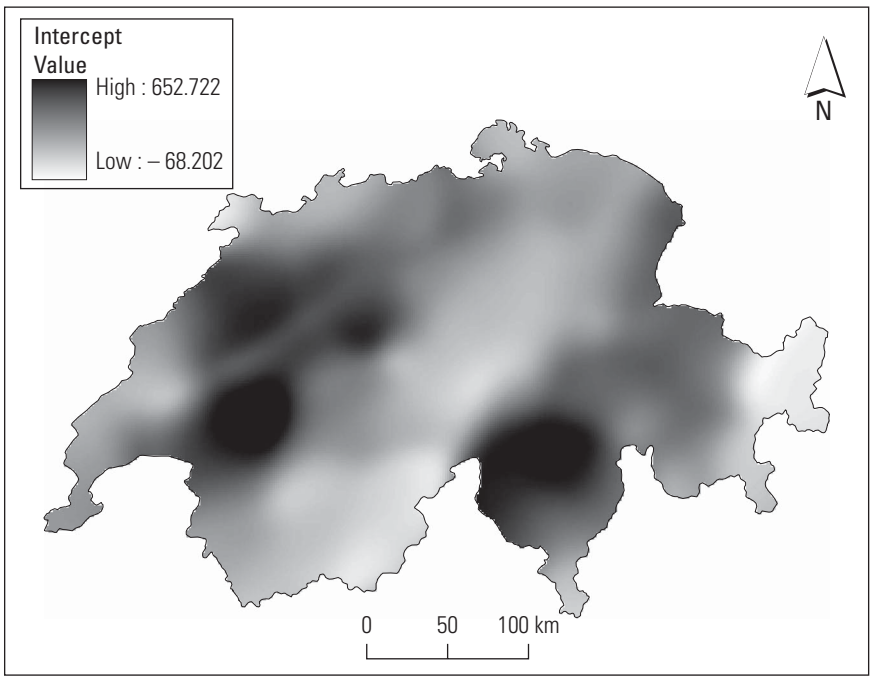

Figure 8.14 GWR intercept: elevation against precipitation amount for 8 May 1986 in Switzerland.

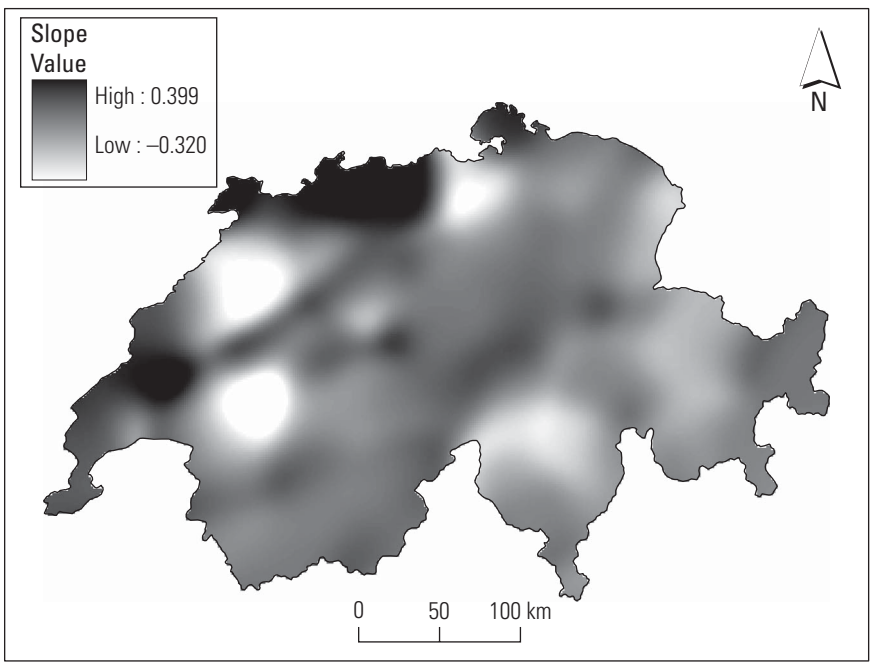

Figure 8.15 GWR slope parameter: elevation against precipitation amount for 8 May 1986 in Switzerland.

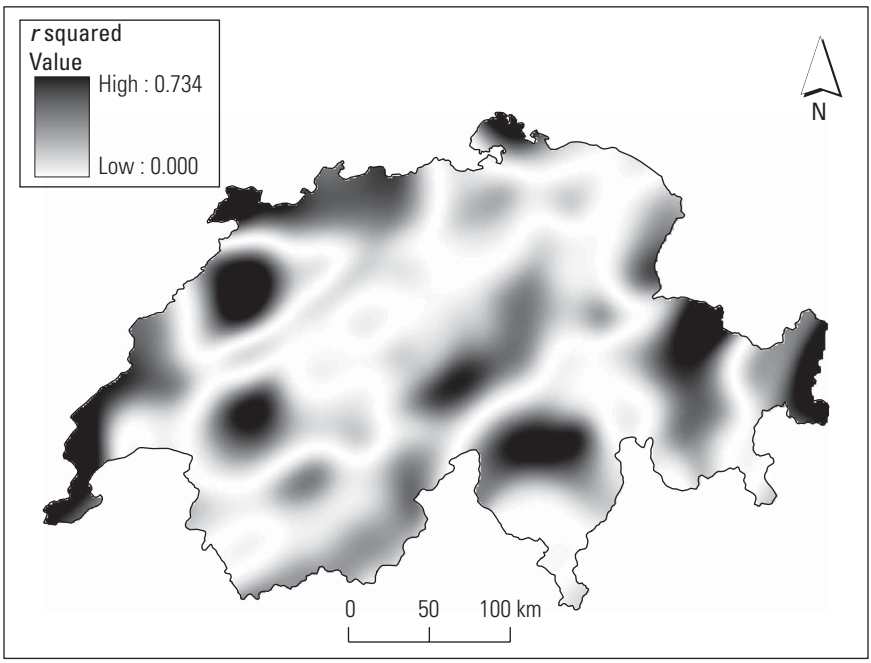

Figure 8.16 GWR coefficient of determination: elevation against precipitation amount for 8 May 1986 in Switzerland.

was used to fit models locally and assess how this relationship varies. The AICc was used to select a GWR kernel bandwidth of 11034.681 m. The GWR intercepts, slopes, and r^2 (as for the Northern Ireland data, the latter was computed outside of GWR 3.0) values are shown in Figures 8.14, 8.15, and 8.16, respectively.

Figures 8.14 and 8.15 provide information about the form of the relationship between elevation and precipitation amount, while Figure 8.16 suggests that the two are strongly related in some regions. Figure 8.15 indicates that the relationship between elevation and precipitation is negative (as for the global regression) or weakly positive in many areas. In some areas, however, there is a strong positive relationship, as indicated by large slope values (Figure 8.15) and large values of the coefficient of determination (r^2; Figure 8.16). The most obvious areas that fit within this category are in the north-west and the far west of Switzerland. The maps could be interpreted further with reference to the case study for Northern Ireland in Section 8.5.3.

Summary

This chapter has introduced approaches for deriving local statistics and for the analysis of spatial autocorrelation at different spatial scales and locally. The principal focus (in terms of space devoted to topics) has been on methods for the analysis of geographically

varying relationships between variables. Knowledge of such approaches opens a wealth of opportunities for exploring spatial data. Exploratory spatial data analysis is a key part of any spatial analysis more generally, and assessing geographical variations in individual variables and in relationships between variables represents a significant improvement on conventional aspatial analyses of geographically referenced data. Many case studies exist that demonstrate some of the possibilities (see, for example, Fotheringham *et al.,* 2002; Lloyd and Shuttleworth, 2005; and Lloyd, 2006).

Further reading

More depth on the methods detailed in this chapter is provided by **Fotheringham *et al.* (2000)**, **Waller and Gotway (2004)**, and **Lloyd (2006)**. The standard account of spatial autocorrelation and its measurement is the book by **Cliff and Ord (1973)**. More on local measures of spatial autocorrelation is provided by **Anselin (1995)** and **Getis and Ord (1996)**. **Fotheringham *et al.* (2002)** provide a detailed account of GWR and the associated software.

➜ The next chapter is concerned with methods for spatial interpolation—that is, methods for prediction of values at unsampled locations.