18 Analysis and modelling of lattice data

Saveliev, A.A., Mukharamova, S.S. and Zuur, A.F.

18.1 Lattice data

In this chapter we consider statistical techniques for analysing spatial units arranged in a lattice pattern. A lattice structure is created when a landscape or region is divided into sub-areas (Cressie 1993). The sub-areas can also be called cells, units or locations. None of the sub-areas can intersect each other, but each shares a boundary edge with one or more of the other sub-areas. An example of a lattice is shown in Figure 18.1. A *regular* lattice is formed if all of the cells have the same form and size. Regular lattices are usually obtained if a region is divided into cells based on a regular grid (e.g., Figure 10.3 for the bird radar data). If a region is divided into cells based on the outlines of natural objects, such as river basins, national boundaries, counties, or postal codes, an *irregular* lattice results. The lattice shown in Figure 18.1 is an example of an irregular lattice.

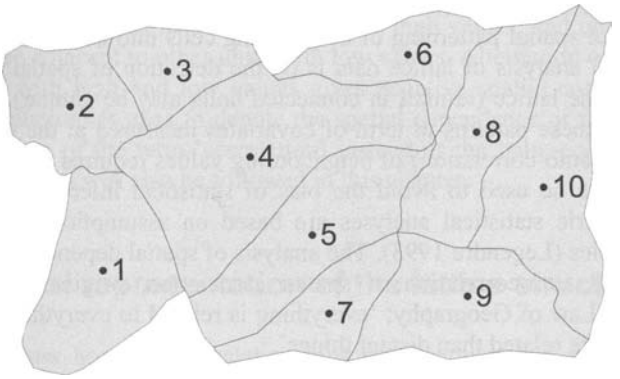

Figure 18.1. Hypothetical irregular lattice. The dots represent the arbitrarily chosen centre of each unit (also called a cell) and the numbers identify each unit.

A typical example of spatial data arranged in a regular lattice pattern is remotely sensed satellite imagery. Satellite data provide spatially distributed information on soil, topography, vegetation, surface temperature, and much more.

These data can also be collected at a variety of spatial resolutions. Image areas can be divided into cells based on a regular grid, and data values for each variable on the grid can be analysed. A wide variety of statistical techniques are used to process satellite imagery. Statistical models based on regular lattices are among the most powerful.

In many biological studies, field data are collected at sites or stations that occur in more irregular configurations than satellite data. The data from these sites are assumed to be representative of the characteristics at the sample unit's position within the lattice. Each unit is usually compared for differences with other sampling sites. Some ecosystem parameters, like regional species richness, however, can only be estimated if the data are aggregated. Aggregation can be done in two ways, namely by (i) using natural spatial units like landscape patches or (ii) using an arbitrary spatial unit like a county or forest inventory stand. Aggregating areas, although sometimes necessary, can cause problems if the phenomenon under study extends beyond the boundaries of the study area. If arbitrary units are used for the analysis, an additional problem called the *modifiable areal unit problem*, can result, in which the size of the spatial unit on which aggregation is applied influences the correlation between the variables. When the size of the cells match the phenomenon under investigation, however, aggregating cells can be a powerful analysis tool, e.g., Openshaw and Taylor (1979), Fotheringham and Wong (1991) or Cressie (1996).

Spatial patterning within the lattice structure of a ecological habitat depends on the interaction of several different forces acting at different spatial scales. These forces range from global climatic factors to local microclimate variations caused by differences in relief and soil characteristics that affect moisture and nutrition availability.

Most of the statistical techniques applied on data that are arranged like a lattice structure take spatial patterning of neighboring cells into account. The purpose of the statistical analysis of lattice data is (i) the detection of spatial patterns in the values over the lattice (rainfall in connected units may be similar) and (ii) an explanation of these patterns in term of covariates measured at the same cells. The dependence (auto-correlation) of neighbouring values requires that special statistical techniques be used to avoid the bias of statistical inference results because most parametric statistical analyses are based on assumptions of independence among samples (Legendre 1993). The analysis of spatial dependence, also referred to as spatial auto-correlation or spatial association, originates from Tobler's (1979) First Law of Geography: 'everything is related to everything else, but near things are more related than distant things'.

Notation

To study spatial associations in data, we assume that the data can be thought of as a random process $\mathbf{Y}$ on a lattice D. The lattice is then a fixed (regular or irregular) collection of spatial objects, and each object has a distinct neighbourhood structure (see Figure 18.1). Mathematically, S denotes the region under study. A_i

denotes the spatial units within S, and S is part of a two-dimensional space $\mathbf{R}^2$. These notations have the following relationships:

$$D = \{A_1, A_2, ..., A_n\} \quad A_i \subset S \subset \mathbf{R}^2 \quad A_1 \cup A_2 \cup ... \cup A_n = S,$$

The first formula defines a lattice consisting of n units (or cells or objects with aerial extent). Within the lattice in Figure 18.1, there are n individual units (A_1 to A_n). The second formula tells us that each unit is contained within the region of interest and the region is part of a two-dimensional space. The third formula designates that the n units together cover the entire study area S. Finally, we need to verify that the lattice consists of units that do not overlap, or that:

$$A_i \cap A_j = \varnothing \text{ for } i \neq j$$

A crucial point, also emphasised in Schabenberger and Pierce (2002), is that all possible units can be enumerated. Hence, n is finite. In Figure 18.1 we can see that none of the units overlap. The random variable at location A_i is defined as $Y_i = Y(A_i)$ and $\mathbf{Y}$ contains the random variables at all n units: $\mathbf{Y} = (Y_1,..,Y_n)$. We denote the observed value for object A_i as y_i and the vector of all n observed values as $\boldsymbol{y} = (y_1,..,y_n)$. So, if we were interested in daily rainfall in Figure 18.1, we have a vector $\boldsymbol{y} = (y_1,..,y_{10})$, and each element represents the measured rainfall in a particular unit. We will use the usual statistical concepts of first- and second-order moments of the variable of interest. These correspond to modelling first-order variation in the mean value $\mu_i = E[Y_i]$ simultaneously with second order variation or spatial dependence between Y_i and Y_j, that is, the covariance between values at two objects (Bailey and Gatrell 1995).

In the analysis of lattice-arranged data, we have to take into account spatial correlation between values at adjacent units. Units with high values (or: above average) are likely to be located near other units with high values; and units with low values should be adjacent to other units with low values. Alternation of neighbouring areal units with high and low values gives negative spatial correlation. The term 'auto-correlation' is used to denote the spatial dependence of the same variable, but the usage of the term 'correlation' instead of the 'auto-correlation' is a common practice that will also be followed in this chapter.

18.2 Numerical representation of the lattice structure

We now discuss how spatial relationships can be quantified using a spatial weight matrix $\mathbf{W}$ in which the elements represent the strength of the spatial structure between the units (Cliff and Ord 1973; Anselin et al. 2004). This spatial weight matrix will be used to calculate the spatial auto-correlation. There are various ways of defining $\mathbf{W}$, and the choice of which one to choose is subjective. All of them are based on the concept of the neighborhood of a unit (A_i). The easiest option is described in Cliff and Ord (1981) and Upton and Fingleton (1985). Construct a binary contiguity matrix (only zeros and ones) by specifying the units that

are adjacent (one), and those that are not (zero). As an example, consider the irregular lattice in Figure 18.1. **W** is of dimension 10-by-10, and its ij^{th} element w_{ij} is given by

$$w_{ij} = \begin{cases} 1 & \text{if the } A_i \text{ and } A_j \text{ have a common border} \\ 0 & \text{otherwise} \end{cases}$$

By definition we have $w_{ii} = 0$. The resulting matrix **W** is in Table 18.1. One of the problems with this definition of the w_{ij}'s is that the common borders between units vary in length. Some areas are only connected by a short border; see for example the pairs (4,8) or (5,6) in Figure 18.1. Cells with such a connection are coloured in grey in Table 18.1. Other units share a longer expanse of border.

Schabenberger and Pierce (2002) used the movements of the chess pieces rook, bishop and queen ('king' would have been more appropriate) to define **W**. Figure 18.2 shows the movements of these pieces. Using the rook movement to define the neighborhood, if A_i is the black unit in the middle, then there are only four other units that have a common border with A_i. The queen movement produces six units with a common border. The matrix **W** in Table 18.1 can be seen as some sort of queen movement for an irregular lattice.

Other options, especially for an irregular lattice, to define **W** are given in Haining (1990) or Schabenberger and Pierce (2002). Using the notation of the latter:

- $w_{ij} = \left\| A_i - A_j \right\|^{-\gamma} \qquad \gamma \geq 0$
- $w_{ij} = \exp(\left\| A_i - A_j \right\|^{-\gamma})$
- $w_{ij} = (l_{ij} / l_i)^{\gamma}$
- $w_{ij} = (l_{ij} / l_i) / \left\| A_i - A_j \right\|^{-\gamma}$

The underlying principle of these definitions is simple; $\| A_i - A_j \|$ defines the spatial separation between units. The closer in space two units A_i and A_j are, the larger the weighting factor w_{ij}. In the last two definitions, l_{ij} is the length of the common border between units A_i and A_j, and l_i is the perimeter of unit A_i. Basically we are defining association between spatial units as a function of physical distance between each unit. Because the distance between regions cannot be uniquely defined, the centre or any other meaningful point can be used. Euclidean distances (Chapter 10) between the i^{th} and j^{th} unit centers are given in Table 18.2.

Table 18.1. Matrix **W** for the objects. An '1' indicates that two areas have a joint border. Empty cells represent zeros (no border in common). Grey cells indicate that the two areas have a short border.

	1	2	3	4	5	6	7	8	9	10
1		1		1						
2	1		1	1						
3		1		1						
4	1	1	1		1	1		1		
5				1		1	1	1	1	
6				1	1			1		
7					1			1	1	
8				1	1	1	1		1	1
9					1		1	1		1
10								1	1	

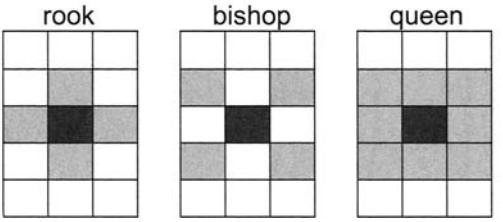

Figure 18.2. Movements of the chess pieces rook, bishop and queen (or actually the king) defining units with a common border for a regular grid.

Table 18.2. Centroid based distance matrix (km). Each value represents the Euclidean distance between the centres of two units. The lower diagonal elements are equal to the upper diagonal elements and were omitted.

	2	3	4	5	6	7	8	9	10
1	5.8	7.2	6.4	7.3	12.9	7.9	13.7	12.6	15.5
2		3.6	6.4	9.4	11.8	11.4	14	15.2	16.5
3			4.1	7.5	8.3	10	10.9	12.9	13.6
4				3.4	6.5	6	7.8	8.9	10.2
5					7	2.8	6.7	5.8	8.1
6						9.3	3.6	8.6	6.5
7							8.1	4.8	8.6
8								5.7	3
9									4.6

After a distance matrix is created using one of the methods above, it is converted into a contiguity matrix. A contiguity matrix is a binary matrix (only zeros and ones) specifying the units that are adjacent (one), and those that are not (zero). In order to do this for Table 18.2, we first need to define a distance cut off point.

Two objects are then considered to be contiguous if their centroids are less than the specified 'cut distance' apart. Figure 18.3 shows the distances from unit 8. The smallest inner circle in Figure 18.3 represents the area that is within 3 km from the centroid of area 8. If we use a 'cut distance' of 3 km, then only unit 10 falls into the neighbourhood of object 8. Hence, for unit 8, we would designate the relationship of the each unit to unit 8 with the following row:

	1	2	3	4	5	6	7	8	9	10
8	0	0	0	0	0	0	0	0	0	1

where '1' designates the only site (10) that is within a radius of 3 kilometres of point 8 and the zeros indicate the units that are outside this interval.

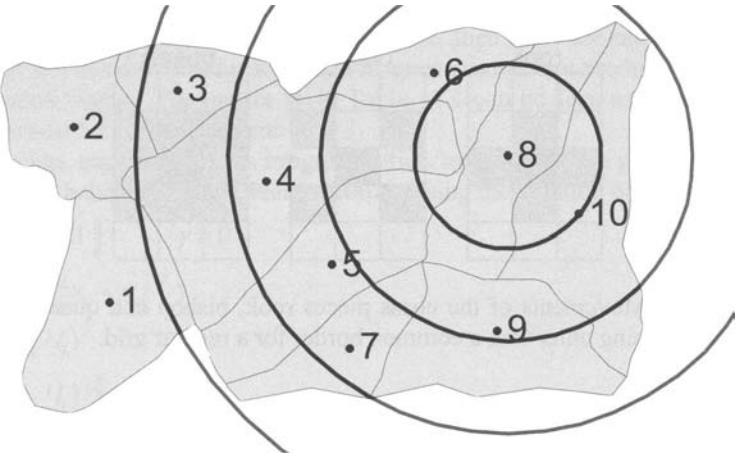

Figure 18.3. Increasing circular neighbourhoods around the object 8. The smallest inner circle has a radius of 3 km and shows that only the centroid of unit 10 is within 3 km of the centroid of unit 8.

Increasing the cut distance to 6, 9 and 12 km, gives the neighbourhoods {6, 9, 10}, {4, 5, 6, 7, 9, 10} and {3, 4, 5, 6, 7, 9, 10}, respectively. So, for the last example, we have:

	1	2	3	4	5	6	7	8	9	10
8	0	0	1	1	1	1	1	0	1	1

This means that only objects 1 and 2 are not within a radius of 12 km of object 8. Hence, we now have a mechanism to convert the distances in Table 18.2 into a contiguity matrix containing only zeros and ones based on the value of a cut distance. Mathematically, the indictor function can be used for this process. An indictor function is a useful mathematical tool of the form:

$$I(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{is } x \text{ is not true} \end{cases}$$

This function can be used as follows. Define $w_{ij}^{(k)}$ as

$$w_{ij}^{(k)} = w_{ij}(d_k, d_{k+1}) = \mathrm{I}(d_k < d_{ij} \leq d_{k+1})$$

where the d_k are the cut distances and the indices i and j refer to centroids. Suppose we use the following cut-distance sequence {0, 3, 6, 9, 12, 15}. For unit $i = 8$, $w_{ij}^{(0)}$ is defined by

$$w_{8j}^{(0)} = w_{8j}(d_0, d_1) = \mathrm{I}(d_0 < d_{8j} \leq d_1) = I(0 < d_{8j} \leq 3)$$

And this gives the same row of zeros and ones as on the previous page:

	1	2	3	4	5	6	7	8	9	10
8	0	0	0	0	0	0	0	0	0	1

The information in the form of zeros and ones tells us which objects are within a 3 km radius of object 8. We can do the same for

$$w_{8,j}(3,6),\ w_{8,j}(6,9),\ w_{8,j}(9,12),\ w_{8,j}(12,15)$$

For example, $w_{ij}^{(1)}$ gives $I(3 < d_{8j} < 6)$, which are all objects that have a distance between 3 and 6 kilometres from object 8. All these terms provide a row of zeros and ones. For $w_{ij}^{(1)}$ we have:

	1	2	3	4	5	6	7	8	9	10
8	0	0	0	0	0	1	0	0	1	0

18.3 Spatial correlation

In previous chapters on time series analysis, we introduced terms like auto- and cross-correlation to quantify the relation within and between time series. In spatial statistics, we use similar tools that allow us to analyse the second-order properties, i.e., the dependence of the variable of interest (e.g., rainfall) between the spatially separated locations. The general formulation is based on a cross-product of the following form (Cliff and Ord 1973; Hubert et al. 1981; Getis 1991):

$$\sum_i \sum_j w_{ij} U_{ij}$$

where U_{ij} is a measure of dissimilarity between the measured variable of interest (e.g., rainfall) at the units i and j and w_{ij} are spatial weights as defined in Section 18.2. We will discuss one such dissimilarity measure in this section, namely Moran's I (Moran 1950) .

The Morans I coefficient

The Moran I coefficient is used to quantify the degree of spatial correlation between neighbouring units. It is defined by

$$I = \frac{n}{w_{++}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{n}{\mathbf{1'W1}} \frac{\mathbf{u'Wu}}{\mathbf{u'u}} \tag{18.1}$$

In this equation, n is the number of units in the lattice, w_{++} is the sum of all weights w_{ij}, y_i is the value of the variable of interest (e.g., rainfall) in object i, and $\bar{y}$ is the mean value of the variable for the whole region. The weights w_{ij} are obtained by any of the methods discussed in Section 18.2. The rightmost part of the equation is just matrix notation for the same thing; **1** is a vector of ones, and **u** contains the centred elements y_i.

The term $w_{ij}(y_i - \bar{y})(y_j - \bar{y})$ is used as a dissimilarity measure (see also the definition of the covariance and correlation coefficients in Chapter 10). In fact, the interpretation is similar to the correlation coefficient. If the values at two adjacent objects are both above average, or below average, then this suggests that there is a positive spatial correlation. If one value is above and the other is below average, a negative correlation is suggested.

Under the null hypothesis of no spatial correlation, the expected value of I is $E[I] = -1/(n-1)$, which is close to 0 for large n. A p-value can be obtained by either assuming asymptotic normality or using a permutation test; see Cliff and Ord (1981), Anselin et al. (2004) or Schabenberger and Pierce (2002) for details. An I value larger than $-1/(n-1)$ means that similar values of the variable of interest, either high or low, are spatially clustered. Negative spatial auto-correlation is harder to interpret.

Example of the Moran I index for tree height data

Figures 18.4A-B show the locations of trees in two 20-by-20 m square plots situated in the Raifa section of the Valga-Kama State reserve (Tatarstan, Russia; see also Chapter 37). Data on the spatial distribution, height, and diameter at breast height (dbh) for five tree species, including *Betula pendula Roth.*, *Acer platanoides L.*, *Tilia cordata Mill.*, *Pinus sylvestris L.* and *Picea × fennica (Regel) Kom.*, were collected by Rogova and co-workers at the Faculty of Ecology, Kazan State University. To create the lattice the Voronoi tessellation[1] (Moller 1994) was used; see Figures 18.4C-D.

[1] The word 'tessellation' means that a particular shape is repeated a large number of times, covering a plane without gaps or overlaps (http://mathforum.org/). Another word for tessellation is tiling, and it is derived from the Greek word *tesseres* which means four. Just think of the square tiles in the bathroom. But just as in the bathroom, the tiles do not have

The Moran I coefficient is used to investigate whether there is any spatial autocorrelation in tree height. The null hypothesis is that there is no spatial correlation. The Moran's I statistic for height of trees for the plot-1-6 data is 0.246. A permutation test on the tree height data indicates that the null hypothesis can be rejected ($p < 0.001$). For plot plot-22-7, the I value is 0.021 ($p = 0.31$) indicating that there is no evidence to reject the null hypothesis. These results suggest, therefore, that plot-1-6 shows spatial correlation of tree heights, but plot-22-7 does not.

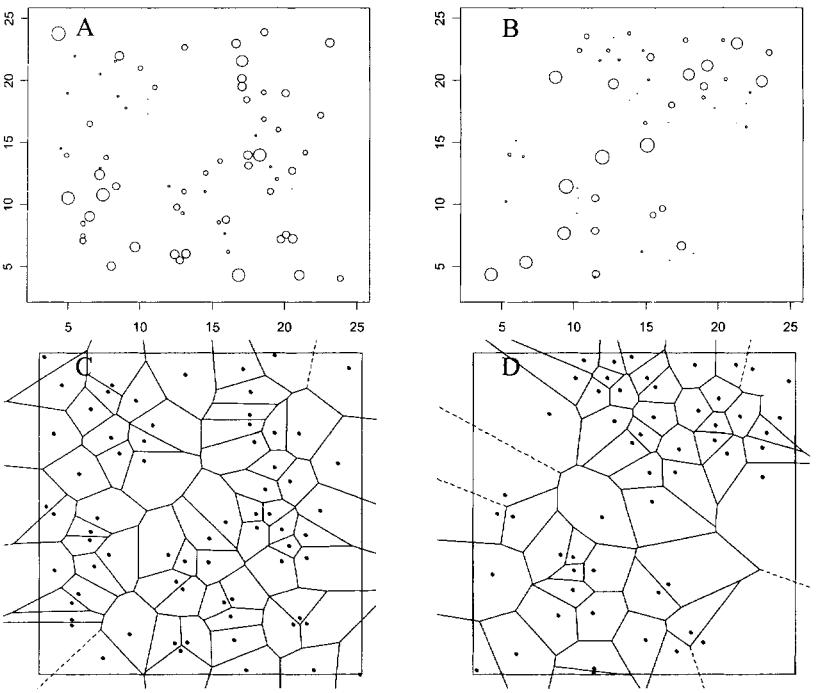

Figure 18.4. Map of the tree locations at the plot-1-6 (A) and plot-22-7 (B). A circle represents the spatial position of a tree, and the diameter of the circles is proportional to tree height. Tree diameter will be used later as an explanatory variable of tree height. C and D: The tree locations were converted into a lattice using so-called Voronoi tessellation. In this process cell border is created; see also Moller (1994).

to be squares. Roman mosaic, for example, can have a huge variety in tessellation patterns. Voronoi diagrams are a special form of tessellation patterns and have been used in many applications. Data other than lattice, e.g., points pattern, can be converted to the lattice using Voronoi tessellation (Moller, 1994). We applied this nearest-neighbourhood tessellation, and as a result the lattice borders are drawn at half the distance between the points.

The spatial correlogram

The Moran I index depends on the choice of how the w_{ij}'s are defined. The w_{ij}'s are zero if the distance is larger than a threshold value or they are equal to one if they are smaller than the threshold. So, choosing the proper threshold value to calculate this index is crucial. To expand the information that we can get from a variable of interest, a range of threshold values can be used. The weights w_{ij} depend on *cutting values*, as defined by the spatial weights $w_{ij}^{(k)}$ (see Section 18.3). We only have to choose a series of cutting distances d_k, d_{k+1}, etc. For each value of k we calculate the Moran statistic. This shows us how the spatial correlation is changing for different distances. The resulting function is called a *spatial correlogram* (Cliff and Ord 1981; Upton and Fingleton 1985). The Moran spatial correlogram is defined by

$$I^{(k)} = \frac{n}{w_{++}^{(k)}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(k)}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{18.2}$$

Note that the only difference between equations (18.1) and (18.2) is the use of the weighting factors. The index $I^{(k)}$ is calculated for different values of k (0, 1, 2,...), and the graphical presentation is a plot of the index k versus $I^{(k)}$. The graph shows how the strength of the dependence changes with distance between units. A permutation test or assymptotic distribution can be used to obtain critical values and p-values (Cliff and Ord 1981; Anselin et al. 2004).

Example of the Moran spatial correlogram for the tree height data

The Moran spatial correlograms for tree height in plot-1-6 and plot-22-7 are shown in Figure 18.5. The cut distance sequences {2.5, 5, 7.5, 10, 12.5, 15} were used. The Moran index is calculated first using only points that are separated by 2.5 m and then it is calculated with points between 2.5 and 5 m, etc. For plot-1-6, , the Moran's $I^{(k)}$ is significantly different from 0 ($p < 0.001$) for the distance band 7.5-10 m; so there is evidence that trees that are separated by 7.5 to 10 m are dependent. For all other classes in plot-1-6 and for plot-22-7, there is no evidence to reject the null hypothesis of no spatial correlation.

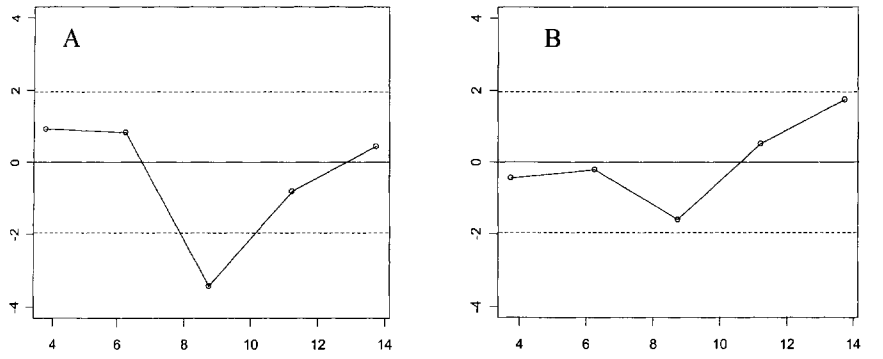

Figure 18.5. Moran's $I^{(k)}$ calculated for trees height at five distance bands (k = 2.5, 5, 7.5, 10, 12.5, 15). A: Plot-1-6. B: Plot-22-7. Expectation (solid line) and two-sided 95% asymptotic critical values (dashed lines) are also given. The y-axis is expressed as a z-score (the difference between $I^{(k)}$ and its mean (over all k) was divided by the standard devition of the $I^{(k)}$s).

18.4 Modelling lattice data

In this section, we will combine regression (and smoothing) models with spatial auto-correlation. Therefore it is essential that you are familiar with the material in Chapters 5 and 7, and knowledge of the time series methods (ARMAX, GLS) in Chapter 16 is also beneficial.

In the previous section, we introduced tools to test whether there is spatial correlation between values at units or cells of lattice-structured data. If the tests indicate that there is indeed spatial auto-correlation, we need to incorporate that correlation within the models that will be applied in the next step of the analysis. Ignoring the spatial auto-correlation structure may cause type I errors and may thus lead to neglecting important exploratory variables and inadequate model selection. The general approach to incorporating auto-correlation into a linear regression model (or any of the extensions discussed in Chapters 6–8) is to model the variable of interest as a function of (i) a systematic part and (ii) a residual component with a covariance structure reflecting fine-scale spatial variation. The systematic part is a phrase used by geostatisticians to express $\boldsymbol{X\beta}$, the effect of explanatory variables. Non-linear effects can easily be incorporated using quadratic terms (Legendre and Legendre 1998) or smoothers (Chapter 7). It is also called the 'spatial trend'. The relevant models are of the form:

$$\text{Response variable} = F(\text{explanatory variables}) + \text{spatial correlated noise}$$

Modelling the systematic spatial trend, i.e. the function F of the explanatory variables, is done using linear regression, generalised linear modelling or generalised additive modelling (Chapters 5 to 7).

In previous chapters, noise within the mathematical model was handled in different ways. In Chapters 5 to 7, noise was assumed to be uncorrelated. In the time series chapters (e.g., chapters 16, 23, 26, and 34), noise was allowed to correlate in time, resulting in generalised least squares (GLS) estimation. To incorporate residual spatial structure within the mathematical models, it is common to assume that neighbouring units have similar values. Some random spatial process is causing the residuals to be spatially correlated.

For lattice data, two popular approaches incorporate residual spatial correlation. These include the *conditional auto-regressive model* (CAR) and the *simultaneously auto-regressive* (SAR) *model*. SAR models are widely used and easy to understand. They are most appropriate for inference studies. CAR models are useful for prediction and spatial interpolation. In this chapter, we will only discuss SAR and their related types. See Cressie (1993) for references and a detailed discussion on both CAR and SAR models. Before we present a discussion of the SAR family of models, however, we need to present the results of a linear regression analysis on the tree height data so that we can refer to these results when presenting the SAR models.

Example of linear regression applied on the tree height data

The following linear regression model was applied on the data from plot-1-6:

$$\text{Tree height}_i = \alpha + \beta\,\text{Tree diameter}_i + \varepsilon_i$$

The numerical output for the linear regression model is as follows:

Variable	Estimate	Std.Err	*t*-value	*p*-value
Intercept	5.820	0.441	13.19	<0.001
Diameter	0.311	0.021	14.45	<0.001

Residual standard error: 2.622 on 68 degrees of freedom
Multiple R-Squared: 0.754, AIC = 337.547
F-statistic: 208.9 on 1 and 68 df, p-value: < 0.001

The interpretation of this type of numerical output was discussed in Chapter 5. The model shows that there is a positive and significant relationship between tree height and diameter. However, the model also assumes that the residuals are independently distributed. To verify this assumption, we applied the Moran's I test on the residuals. We can do this because we have access to the spatial coordinates of the residuals (Figure 18.4). The result was $I = 0.115$ ($p = 0.026$), which means that there is evidence of spatial auto-correlation in the residuals. The Moran's I coefficient for this specific situation is given by

$$I = (n/\mathbf{1}'\mathbf{W}\mathbf{1}) \times (\mathbf{e}'\mathbf{W}\mathbf{e}/(\mathbf{e}'\mathbf{e})) \tag{18.3}$$

This is a similar matrix notation as in equation (18.1). The residuals $\mathbf{e}$ are from the linear regression of tree height on the explanatory variable tree diameter. The

matrix **W** contains the weights (Section 18.2), and details on the distribution of the statistic can be found in Cliff and Ord (1981).

If the test indicates that the spatial error auto-correlation is significant, we violate the underlying assumption of linear regression that requires independent residuals. This means that we cannot trust the p-values for the t- and F-statistics in the above regression analysis. One option to lessen spatial auto-correlation is to try to include more explanatory variables in the model. The residual pattern may be present because an important explanatory variable is missing from the analysis. Alternatively, the SAR model can be applied, which is described next.

Simultaneous auto-regressive model

At this point, you may want to read the time series chapter (Chapter 16), which contains relevant key terminology and notation applicable to the SAR models. In time series, the auto-regressive time series model (AR) of order p is given by

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t$$

The α and β_js are unknown regression parameters, and the errors ε_i are independent and normally distributed with the zero mean and variance σ^2. So, the explanatory variables in an AR time series model are lagged response variables. For spatial data, we consider the spatial weights of units in the lattice to construct regression models that take into account spatial auto-correlation. The SAR model is defined similarly (Ord 1975):

$$Y_i = \mu_i + \rho \sum_j w_{ij}(Y_j - \mu_j) + \varepsilon_i \tag{18.4}$$

The w_{ij} are spatial weights (Section 18.2), Y_i is a random variable corresponding to the i^{th} unit, $E[Y_i] = \mu_i$, ρ is a model parameter, and the errors ε_i are assumed to be independent and normally distributed with the zero mean and variance σ^2. Note the similarity between the AR time series model and the SAR model. Both contain lagged response variables as explanatory variables, and both have the same assumptions on the error term. This explains why we call the model in equation (18.4) a simultaneous auto-regressive model. The parameter ρ measures the strength of the spatial correlation. It is common to assume that $\mu_i = \alpha + \mathbf{X}_i\boldsymbol{\beta}$, and it represents the F(explanatory variables) component mentioned above. The intercept can be included in **X** and $\boldsymbol{\beta}$ by using a column with only ones in **X**. As a result we have $E[Y_i] = \mathbf{X}_i\boldsymbol{\beta}$. In matrix notation, the SAR model in equation (18.4) can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \quad \text{and} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{18.5}$$

W contains the weights and **Y**, $\boldsymbol{\varepsilon}$ are n-by-1 vectors, and **I** is the identity matrix. **X** is the n-by-m matrix containing explanatory variables and $\boldsymbol{\beta}$ is an m-by-1 vector of the regression coefficients. A common approach is to use the spatial coordinates (e.g., latitude and longitude) as explanatory variables in **X**. Details for pa-

rameter estimations can be found in Ord (1975) and Anselin (1988). Standard statistics (pseudo-R^2 and AIC, BIC) for the model, goodness-of-fit and confidence intervals for parameters can be calculated, and parameter significance estimation is based on the asymptotic normality (Ord 1975).

Example of a SAR model for the tree height data

A SAR model was applied on the plot-1-6 tree height data with the same explanatory variables that were used to create the linear model. The following model was applied:

$$Height_i = \alpha + \beta Diameter_i + \rho \sum_j w_{ij}(Height_j - \beta Diameter_i) + \varepsilon_i$$

The numerical output from the fitted SAR model is given below:

Variable	Estimate	Std.Error	*z*-value	*p*-value
Intercept	5.144	1.107	4.645	<0.001
Diameter	0.304	0.023	13.022	<0.001
rho	0.074	0.113	0.660	0.509

Residual standard error: 2.574
AIC: 339.06
LM test for residual auto-correlation test value: 1.553 (p=0.212)

Although adding the auto-regression parameter ρ to the model removes the residual auto-correlation (Moran's I coefficient was calculated for the residuals), ρ is not significantly different from 0 (p = 0.509). Also note that the residual standard error is only slightly lower than the residual of the linear regression model (2.574 against 2.622), but the AIC for the SAR model is slightly higher than the linear regression model (339.06 against 337.55). Just as in linear regression, the model with the lowest AIC is preferred.

18.5 More exotic models

We now discuss a series of models that are closely related to SAR models. They differ from the SAR model discussed above mainly in how the residual error is incorporated into the model.

Spatial moving average model

When we introduced time series techniques in Chapter 16, we first introduced AR models, then moving average (MA) models, and then combined the AR and MA models into ARMAX models. Recall that the MA time series model was given by

$$Y_t = \alpha + \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \gamma_2 \varepsilon_{t-2} + \ldots + \gamma_p \varepsilon_{t-p}$$

The variable Y_t is modelled in terms of current and past error terms with unknown regression coefficients γ_i. We can do something similar for the spatial models. The variable of interest is modelled as a function of explanatory variables and residual patterns from different units, and not surprisingly the resulting model is called a spatial moving average model (SMA). The model is given by (Ord 1975):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{1i} + \ldots + \beta_m X_{mi} + u_i \\ u_i &= \lambda \sum_j w_{ij} u_j + \varepsilon_i \end{aligned} \tag{18.6}$$

The regression parameters β_i model the effect of the explanatory variables on the response variable Y_i, u_i is spatially correlated noise and λ is the error autoregression coefficient. If λ is zero, there is no spatial correlation and u_i becomes independently distributed noise. In matrix notation, equation (18.6) becomes

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{U} \qquad \mathbf{U} = \lambda \mathbf{WU} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Y contains the data, and **U** is a vector of the spatially correlated errors.

Example of a SMA model for the tree height data

The following SMA model was applied on the plot-1-6 tree height data:

$$\begin{aligned} Height_i &= \alpha + \beta Diameter_i + u_i \\ u_i &= \lambda \sum_j w_{ij} u_j + \varepsilon_i \end{aligned} \tag{18.7}$$

The numerical output of the SMA model is given below:

Variable	Estimate	Std.Error	*z*-value	*p*-value
Intercept	5.806	0.520	11.147	<0.001
Diameter	0.310	0.022	14.006	<0.001
Lambda	0.265	0.175	1.521	0.128

Residual standard error: 2.523
AIC: 337.34

Although adding the nuisance parameter λ to the model gives minor improvement in the standard error (2.523 against the 2.622 for the linear regression model) and AIC (337.34 against 337.55 for the linear regression), it is not significant (*p*-value 0.128). The AICs for the linear regression, SAR and SMA model are, respectively, 337.55, 339.06 and 337.34, which indicates that the SAR model is less adequate in describing tree height than the SMA model, although the significance of the nuisance parameter λ in the SMA model is low ($p = 0.128$).

As mentioned, spatial auto-correlation in the models may be caused by not including an important explanatory variable in the model. For the tree data, we have ignored the species information of the tree (there are four different species). Including 'species' as a nominal explanatory variable to the linear regression model gives:

$$Height_i = \alpha + \beta Diameter_i + factor(Species_i) + \varepsilon_i \tag{18.8}$$

The numerical output shows that adding the nominal variable gives a better overall fit of the model to the data:

Variable	Estimate	Std.Error	*t*-value	*p*-value
Intercept	3.058	1.673	1.828	0.072
Diameter	0.320	0.030	10.423	<0.001
Species2	2.262	1.605	1.409	0.163
Species3	3.910	1.311	2.982	0.004
Species4	0.790	1.792	0.441	0.660

Residual standard error: 2.354 on 65 degrees of freedom
Multiple *R*-Squared: 0.8107
F-statistic: 69.61 on 4 and 65 df, *p*-value: < 2.2e-16
AIC = 325.317

Note that the AIC is now 325.317, which is considerably lower compared with the model without the nominal variable species. Again using the Moran's I test for spatial correlation of the residuals gave $I = 0.041$ ($p = 0.18$) and there was no evidence to reject the null hypothesis of no spatial auto-correlation. At this point we could stop and consider the model in equation (18.8) as the most optimal model. But purely for curiosity, we also applied the SMA equivalent of the model in equation (18.8):

$$\begin{aligned} Height_i &= \alpha + \beta Diameter_i + factor(Species_i) + u_i \\ u_i &= \lambda \sum_j w_{ij} u_j + \varepsilon_i \end{aligned} \tag{18.9}$$

The results are as follows:

Variable	Estimate	Std.Error	*z*-value	*p*-value
Intercept	2.827	1.611	1.754	0.079
Diameter	0.325	0.029	10.862	<0.001
Species2	2.441	1.531	1.594	0.110
Species3	3.987	1.261	3.159	0.001
Species4	1.185	1.735	0.683	0.494
Lambda	0.156	0.186	0.841	0.401

Residual standard error: 2.253
AIC: 326.67

The nuisance parameter λ is not significant ($p = 0.401$), and although the residuals standard error has slightly decreased, the AIC favours the linear regression model with species included as a nominal explanatory variable.

Locally linear spatial models

Recall that in Chapter 17 we defined a regression model in which the regression parameters were allowed to change over time (dynamic regression models). We can formulate a similar model for spatial data. In the linear regression model $\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, let the regression coefficients $\boldsymbol{\beta}$ be a function of the spatial location. One example of such an approach is when the regression coefficients are linear functions of the additional variables, including the spatial coordinates (Casetti and Emilio 1972):

$$\beta_j = \gamma_{j,0} + \gamma_{j,1} z_1 + \ldots + \gamma_{j,k} z_k$$

where $j = 1,..,m$ and m is the number of exploratory variables. For example, if z_1 and z_2 are longitude and latitude coordinates, the regression coefficients can be linear functions of the form

$$\beta_j = \gamma_{j,0} + \gamma_{j,1}\text{Longtitude} + \gamma_{j,2}\text{Latitude}$$

Examples of such models can be found in Jones and Casetti (1992), and details of the parameter estimation method are in Casetti (1982). Such a model for the tree example would be of the form (results are not presented here):

$$\text{Height}_i = \alpha + \beta\ \text{Diameter}_i + \varepsilon_i$$

$$\beta = \gamma_1\ \text{Latitude} + \gamma_2\ \text{Longitude}$$

Linear regression model with correlated errors — LM(ce)

The last model we want to introduce is the linear regression model with correlated errors, denoted by LM(e). As with the spatial moving average, we can link the LM(ce) process to the time series models. Recall that for the dynamic factor analysis, we used a symmetric non-diagonal error covariance matrix. The same is done here. The LM(ce) model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \text{and} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}),$$

Technically, the error covariance matrix is modelled as symmetric and positive-definite. The unknown parameters are the regression parameters $\boldsymbol{\beta}$ and the elements in $\boldsymbol{\Lambda}$. Using matrix algebra, it is easy to rewrite the LM(ce) model to

$$\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*, \quad \text{where} \quad \boldsymbol{\varepsilon}^* \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

and $\mathbf{Y}^{*}=\mathbf{\Lambda}^{-0.5}\mathbf{Y}$, $\boldsymbol{\varepsilon}^{*}=\mathbf{\Lambda}^{-0.5}\boldsymbol{\varepsilon}$ and $\mathbf{X}=\mathbf{\Lambda}^{-0.5}\mathbf{X}$. As this is again the linear regression model with uncorrelated residuals (note the identity matrix $\mathbf{I}$ in the normal distribution), the ordinary least squares (Chapter 5) can be used to estimate the parameters. However, alternative estimation routines are also available (Amemiya 1985; Greene 2000). Just as in Chapter 16, one has to choose the structure of $\mathbf{\Lambda}$. This modelling approach is further discussed in Chapter 37 and presented with a detailed example of the modelling approach..

18.6 Summary

The process of analysing spatial data can follow many pathways. Figure 18.6 shows the decision processes to analyse spatial data that have either a regular or irregular lattice structure. In its most simple form, spatial analysis is modelled using linear regression. If a linear regression model is applied, the residuals are assumed to be independent. The Morgan *I* index is used to test whether the spatial residuals from the linear regression are really independent. With luck, the test give no reason to reject the null hypothesis that there is no residual spatial auto-correlation. If there is spatial residual correlation, several options exist. More covariates or interaction terms can be added to the regression model. Smoothing methods can be applied. If simple procedures do not create an appropriate model, then alternative models exist that are specially designed to take spatial dependence into account. Choosing the appropriate spatial model, whether it is a SAR, SMA, local linear spatial model, or LM(ce), will depend on the data structure, the underlying questions, the residual patterns, and the AIC results computed for each model. An example of these methods is given in Chapter 37.

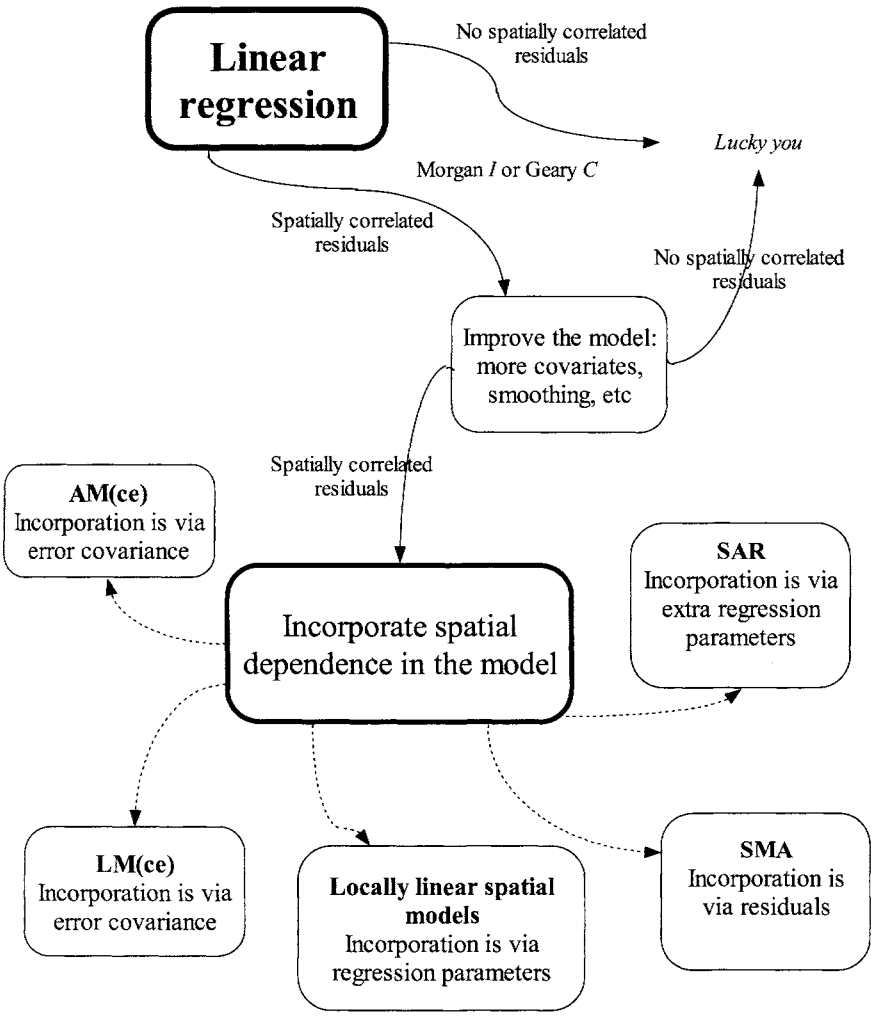

Figure 18.6. Flowchart showing the decision process for the modelling of data measured on a lattice. If a linear regression model is applied, the residuals are assumed to be independent. The Morgan I index can be used to test for spatial independence in the residuals. If you are lucky, there is no evidence to reject the null hypothesis of no spatial correlation. If there is spatial correlation, it is time for action. Either add more covariates or interaction terms, using smoothing methods, and if this does not help, consider one of the many alternative models that take into account spatial dependence. The choice, which model to choose, depends on the data, underlying questions, and residual patterns, AIC, etc. of each model. AM(ce) represents the smoothing equivalen of the LM(ce) model. 'Extra regression parameters' in SAR refers to using lagged response variables.