Chapter 4

Basics of areal data models

We now present a development of exploratory tools and modeling approaches that are customarily applied to data collected for areal units. Again, this literature is sometimes referred to as discrete spatial modeling to reflect the fact that we are only specifying a joint model for a finite set of random variables. We have in mind general, possibly irregular geographic units, but of course include the special case of regular grids of cells (pixels). Indeed, many of the ensuing models have been proposed for regular lattices of points and parameters, and sometimes even for point-referenced data (see Chapter 12 on the problem of inverting very large matrices).

In the context of areal units the general inferential issues are the following:

(i) Is there spatial pattern? If so, how strong is it? Intuitively, "spatial pattern" suggests that measurements for areal units which are near to each other will tend to take more similar values than those for units far from each other. Though you might "know it when you see it," this notion is evidently vague and in need of quantification. Indeed, with independent measurements for the units, we expect to see *no pattern*, i.e., a completely random arrangement of larger and smaller values. But again, randomness will inevitably produce some patches of similar values.

(ii) Do we want to smooth the data? If so, how much? Suppose, for example, that the measurement for each areal unit is a count, say, a number of cancers. Even if the counts were independent, and perhaps even after population adjustment, there would still be extreme values, as in any sample. Are the observed high counts more elevated than would be expected by chance? If we sought to present a surface of expected counts we might naturally expect that the high values would tend to be pulled down, the low values to be pushed up. This is the notion of smoothing. No smoothing would present a display using simply the observed counts. Maximal smoothing would result in a single common value for all units, clearly excessive. Suitable smoothing would fall somewhere in between, and take the spatial arrangement of the units into account.
Of course, how much smoothing is appropriate is not readily defined. In particular, for model-based smoothers such as we describe below, it is not evident what the extent of smoothing is, or how to control it. Specification of a utility function for smoothing (as attempted in Stern and Cressie, 1999) would help to address these questions but does not seem to be considered in practice..

(iii) For a new areal unit or set of units, how can we infer about what data values we expect to be associated with these units? That is, if we modify the areal units to new units, e.g., from zip codes to census block groups, what can we say about the cancer counts we expect for the latter, given those for the former? This is the so-called *modifiable areal unit problem (MAUP)*, which historically (and in most GIS software packages) is handled by crude areal allocation. Sections 7.2 and 7.3 propose model-based methodology for handling this problem.

As a matter of fact, in order to facilitate interpretation and better assess uncertainty, we will suggest model-based approaches to treat the above issues, as opposed to the more descriptive or algorithmic methods that have dominated the literature and are by now widely available in GIS software packages. We will also introduce further flexibility into these models by examining them in the context of regression. That is, we assume interest in explaining the areal unit responses and that we have available potential covariates to do this. These covariates may be available at the same or at different scales from the responses, but, regardless, we will now question whether there remains any spatial structure adjusted for these explanatory variables. This suggests that we may not try to model the data in a spatial way directly, but instead introduce spatial association through random effects. This will lead to versions of generalized linear mixed models (Breslow and Clayton, 1993). We will generally view such models in the hierarchical fashion that is the primary theme of this text.

4.1 Exploratory approaches for areal data

We begin with the presentation of some tools that can be useful in the initial exploration of areal unit data. The primary concept here is a *proximity matrix*, W. Given measurements $Y_1, \ldots, Y_n$ associated with areal units $1, 2, \ldots, n$, the entries w_{ij} in W spatially connect units i and j in some fashion. (Customarily w_{ii} is set to 0.) Possibilities include binary choices, i.e., $w_{ij} = 1$ if i and j share some common boundary, perhaps a vertex (as in a regular grid). Alternatively, w_{ij} could reflect "distance" between units, e.g., a decreasing function of intercentroidal distance between the units (as in a county or other regional map). But distance can be returned to a binary determination. For example, we could set $w_{ij} = 1$ for all i and j within a specified distance. Or, for a given i, we could get $w_{ij} = 1$ if j is one of the K nearest (in distance) neighbors of i. The preceding choices suggest that W would be symmetric. However, for irregular areal units, this last example provides a setting where this need not be the case. Also, the w_{ij}'s may be standardized by $\sum_j w_{ij} = w_{i+}$. If $\widetilde{W}$ has entries $\widetilde{w}_{ij} = w_{ij}/w_{i+}$, then evidently $\widetilde{W}$ is row stochastic, i.e., $\widetilde{W}\mathbf{1} = \mathbf{1}$, but now $\widetilde{W}$ need not be symmetric.

As the notation suggests, the entries in W can be viewed as weights. More weight will be associated with j's closer (in some sense) to i than those farther away from i. In this exploratory context (but, as we shall see, more generally) W provides the mechanism for introducing spatial structure into our formal modeling.

Lastly, working with distance suggests that we can define distance bins, say, $(0, d_1], (d_1, d_2], (d_2, d_3]$, and so on. This enables the notion of *first-order neighbors* of unit i, i.e., all units within distance d_1 of i, *second-order neighbors*, i.e., all units more than d_1 but at most d_2 from i, *third-order neighbors*, and so on. Analogous to W we can define $W^{(1)}$ as the proximity matrix for first-order neighbors. That is, $w_{ij}^{(1)} = 1$ if i and j are first-order neighbors, and equal to 0 otherwise. Similarly we define $W^{(2)}$ as the proximity matrix for second-order neighbors; $w_{ij}^{(2)} = 1$ if i and j are second-order neighbors, and 0 otherwise, and so on to create $W^{(3)}$, $W^{(4)}$, etc.

Of course, the most obvious exploratory data analysis tool for lattice data is a map of the data values. Figure 4.1 gives the statewide average verbal SAT exam scores as reported by the College Board and initially analyzed by Wall (2004). Clearly these data exhibit strong spatial pattern, with midwestern states and Utah performing best, and coastal states and Indiana performing less well. Of course, before jumping to conclusions, we must realize there are any number of spatial covariates that may help to explain this pattern; for instance, the percentage of eligible students taking the exam (Midwestern colleges have historically relied on the ACT exam, not the SAT, and only the best and brightest students in these states typically take the latter exam). Still, the map of these raw data shows significant spatial pattern.

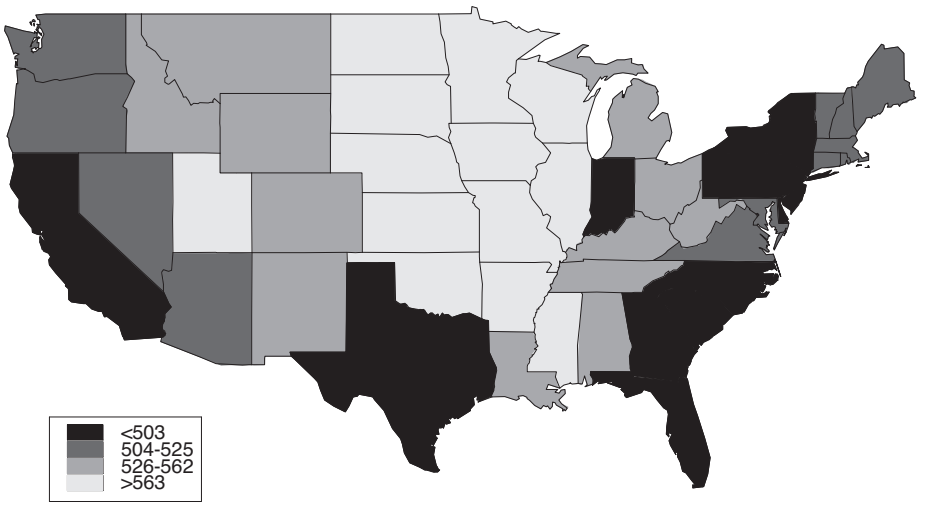

Figure 4.1 *Choropleth map of 1999 average verbal SAT scores, lower 48 U.S. states and the district of Columbia.*

4.1.1 Measures of spatial association

Two standard statistics that are used to measure strength of spatial association among areal units are Moran's I and Geary's C (see, e.g., Ripley, 1981, Sec. 5.4). These are spatial analogues of statistics for measuring association in time series, the lagged autocorrelation coefficient and the Durbin-Watson statistic, respectively. They can also be seen to be areal unit analogues of the empirical estimates for the correlation function and the variogram, respectively. Recall that, for point-referenced data, the empirical covariance function (2.12) and semivariogram (2.9), respectively, provide customary nonparametric estimates of these measures of association.

Moran's I takes the form

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (Y_i - \overline{Y})^2} . \tag{4.1}$$

I is not strictly supported on the interval $[-1, 1]$. It is evidently a ratio of quadratic forms in $\mathbf{Y}$, which provides the idea for obtaining approximate first and second moments through the delta method (see, e.g., Agresti, 2002, Ch. 14). Moran shows under the null model where the Y_i are i.i.d., I is asymptotically normally distributed with mean $-1/(n-1)$ and a rather unattractive variance of the form

$$\mathrm{Var}(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2} . \tag{4.2}$$

In (4.2), $S_0 = \sum_{i \neq j} w_{ij}$, $S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2$, and $S_2 = \sum_k (\sum_j w_{kj} + \sum_i w_{ik})^2$. We recommend the use of Moran's I as an exploratory measure of spatial association, rather than as a "test of spatial significance."

For the data mapped in Figure 4.1, we used the `moran.test` function in the `spdep` package in `R` (see Section 4.5.2) to obtain a value for Moran's I of 0.6125, a reasonably large value. The associated standard error estimate of 0.0979 suggests very strong evidence against the null hypothesis of no spatial correlation in these data.

Geary's C takes the form

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2\left(\sum_{i \neq j} w_{ij}\right)\sum_i (Y_i - \overline{Y})^2} \,. \tag{4.3}$$

C is never negative, and has mean 1 for the null model; *small* values (i.e., between 0 and 1) indicate *positive* spatial association. Also, C is a ratio of quadratic forms in $\mathbf{Y}$ and, like I, is asymptotically normal if the Y_i are i.i.d. We omit details of the distribution theory, recommending the interested reader to Cliff and Ord (1973), or Ripley (1981, p. 99).

Using the `geary.test` function on the SAT verbal data in Figure 4.1, we obtained a value of 0.3577 for Geary's C, with an associated standard error estimate of 0.0984. Again, the marked departure from the mean of 1 indicates strong positive spatial correlation in the data.

Convergence to asymptotic normality for a ration of quadratic forms is extremely slow. We may believe the significant rejection of independence using the asymptotic theory for the example above because the results are so extreme. However, if one truly seeks to run a significance test using (4.1) or (4.3), our recommendation is a Monte Carlo approach. Under the null model the distribution of I (or C) is invariant to permutation of the Y_i's. The exact null distribution of I (or C) requires computing its value under all $n!$ permutations of the Y_i's, infeasible for n in practice. However, a Monte Carlo sample of say 1000 permutations, including the observed one, will position the observed I (or C) relative to the remaining 999, to determine whether it is extreme (perhaps via an empirical p-value). Again using `spatial.cor` function on our SAT verbal data, we obtained empirical p-values of 0 using both Moran's I and Geary's C; *no* random permutation achieved I or C scores as extreme as those obtained for the actual data itself.

A further display that can be created in this spirit is the *correlogram*. Working with say, I, in (4.1) we can replace w_{ij} with the previously defined $w_{ij}^{(1)}$ and compute, say $I^{(1)}$. Similarly, we can replace w_{ij} with $w_{ij}^{(2)}$ and obtain $I^{(2)}$. A plot of $I^{(r)}$ vs. r is called a correlogram and, if spatial pattern is present, is expected to decline in r initially and then perhaps vary about 0. Evidently, this display is a spatial analogue of a temporal lag autocorrelation plot (e.g., see Carlin and Louis, 2000, p. 181). In practice, the correlogram tends to be very erratic and its information context is often not clear.

With large, regular grids of cells as we often obtain from remotely sensed imagery, it may be of interest to study spatial association in a particular direction (e.g., east-west, north-south, southwest-northeast, etc.). Now the spatial component reduces to one dimension and we can compute lagged autocorrelations (lagged appropriately to the size of the grid cells) in the specific direction. An analogue of this was proposed for the case where the Y_i are binary responses (e.g., presence or absence of forest in the cell) by Agarwal, Gelfand, and Silander (2002). In particular, Figure 4.2 shows rasterized maps of binary land use classifications for roughly 25,000 1 km $\times$ 1 km pixels in eastern Madagascar; see Agarwal et al. (2002) as well as Section 7.5 for further discussion.

While the binary map in Figure 4.2 shows spatial pattern in land use, we develop an additional display to provide quantification. For data on a regular grid or lattice, we calculate binary analogues of the sample autocovariances, using the 1 km $\times$ 1 km resolution with four illustrative directions: East (E), Northeast (NE), North (N), and Northwest (NW). In particular for any pair of pixels, we can identify, say, a Euclidean distance and direction between them by labeling one as X and the other as Y, creating a correlated binary pair. Then, we can go to the lattice and identify all pairs which share the same distance and direction. The collection of all such (X,Y) pairs yields a 2×2 table of counts (with table cells labeled as X=0, Y=0; X=0, Y=1; X=1, Y=0, X=1, Y=1). The resultant log-odds ratio measures the association between pairs in that direction at that distance. (Note that

Figure 4.2 *Rasterized north and south regions (1 km × 1 km) with binary land use classification overlaid.*

if we followed the same procedure but reversed direction, e.g., changed from E to W, the corresponding log odds ratio would be unchanged.)

In Figure 4.3, we plot log odds ratio against direction for each of the four directions. Note that the spatial association is quite strong, requiring a distance of at least 40 km before it drops to essentially 0. This suggests that we would not lose much spatial information if we work with the lower (4 km × 4 km) resolution. In exchange we obtain a richer response variable (17 ordered levels, indicating number of forested cells from 0 to 16) and a substantial reduction in number of pixels (from 26,432 to 1, 652 in the north region, from 24,544 to 1, 534 in the south region) to facilitate model fitting.

4.1.2 Spatial smoothers

Recall from the beginning of this chapter that often a goal for, say, a choropleth map of the Y_i's is *smoothing*. Depending upon the number of classes used to make the map, there is already some implicit smoothing in such a display (although this is not *spatial* smoothing, of course).

The W matrix directly provides a spatial smoother; that is, we can replace Y_i by $\widehat{Y}_i = \sum_j w_{ij} Y_j / w_{i+}$. This ensures that the value for areal unit i "looks like" its neighbors, and that the more neighbors we use in computing $\widehat{Y}_i$, the more smoothing we will achieve. In fact, $\widehat{Y}_i$ may be viewed as an unusual smoother in that it ignores the value actually observed for unit i. As such, we might revise the smoother to

$$\widehat{Y}_i^* = (1-\alpha) Y_i + \alpha \widehat{Y}_i \,, \tag{4.4}$$

where $\alpha \in (0, 1)$. Working in an exploratory mode, various choices may be tried for α, but for any of these, (4.4) is a familiar *shrinkage* form. Thus, under a specific model with a suitable loss function, an optimal α could be sought. Finally, the form (4.4), viewed generally as

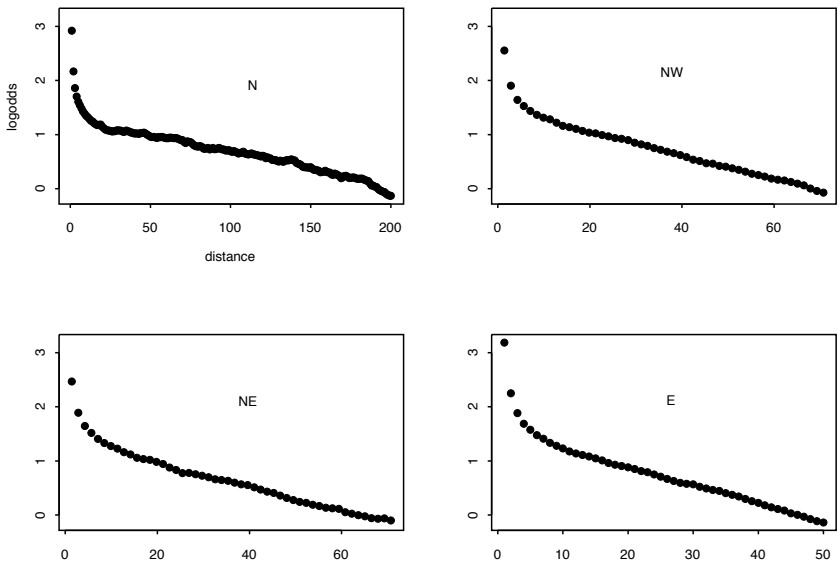

Figure 4.3 *Land use log-odds ratio versus distance in four directions.*

a linear combination of the Y_j, is customarily referred to as a *filter* in the GIS literature. In fact, such software will typically provide choices of filters, and even a default filter to automatically smooth maps.

In Section 5.1 we will present a general discussion revealing how smoothing (shrinkage) emerges as a byproduct of the hierarchical models we propose to use to explain the Y_i. In particular, when W is used in conjunction with a stochastic model (as in Section 4.3), the $\widehat{Y}_i$ are updated across i and across Monte Carlo iterations as well. So the observed Y_i will affect the eventual $\widehat{Y}_i$; we achieve model-driven smoothing and a "manual" inclusion of Y_i as in (4.4) with some choice of α is unnecessary.

4.2 Brook's Lemma and Markov random fields

A useful technical result for obtaining the joint distribution of the Y_i in some of the models we discuss below is *Brook's Lemma* (Brook, 1964). The usefulness of this lemma is exposed in Besag's (1974) seminal paper on conditionally autoregressive models.

It is clear that given $p(y_1, \ldots, y_n)$, the so-called *full conditional* distributions, $p(y_i|y_j, j \neq i)$, $i = 1, \ldots, n$, are uniquely determined. Brook's Lemma demonstrates the converse and, in fact, enables us to constructively retrieve the unique joint distribution determined by these full conditionals. But first, it is also clear that we cannot write down an arbitrary set of full conditional distributions and assert that they uniquely determine the joint distribution. To see this, let $Y_1|Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, \sigma_1^2)$ and let $Y_2|Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, \sigma_2^2)$, where N denotes the normal (Gaussian) distribution. Intuitively, it seems that a mean for Y_1 given Y_2 which is linear in Y_2 is incompatible with a mean for Y_2 given Y_1 which is linear in Y_1^3. More formally, we see that

$$E(Y_1) = E[E(Y_1|Y_2)] = E[\alpha_0 + \alpha_1 Y_2] = \alpha_0 + \alpha_1 E(Y_2) \,, \tag{4.5}$$

i.e., $E(Y_1)$ and $E(Y_2)$ are linearly related. But in fact, it must also be the case that

$$E(Y_2) = E[E(Y_2|Y_1)] = E[\beta_0 + \beta_1 Y_1] = \beta_0 + \beta_1 E(Y_1^3) \,. \tag{4.6}$$

Equations (4.5) and (4.6) could simultaneously hold only in trivial cases, so the two mean specifications are *incompatible*. Thus we can say that $f(y_1|y_2)$ and $f(y_2|y_1)$ are incompatible with regard to determining $p(y_1, y_2)$. We do not examine conditions for compatibility of conditional distributions here, although there has been considerable work in this area (see, e.g., Arnold and Strauss, 1991, and references therein).

Another point is that $p(y_1 \ldots, y_n)$ may be improper even if $p(y_i|y_j, j \neq i)$ is proper for all i. As an elementary illustration, consider $p(y_1, y_2) \propto \exp[-\frac{1}{2}(y_1 - y_2)^2]$. Evidently $p(y_1|y_2)$ is $N(y_2, 1)$ and $p(y_2|y_1)$ is $N(y_1, 1)$, but $p(y_1, y_2)$ is improper. Casella and George (1992) provide a similar example in a bivariate exponential (instead of normal) setting.

Brook's Lemma notes that

$$\begin{aligned} p(y_1, \ldots, y_n) &= \frac{p(y_1|y_2, \ldots, y_n)}{p(y_{10}|y_2, \ldots, y_n)} \cdot \frac{p(y_2|y_{10}, y_3, \ldots, y_n)}{p(y_{20}|y_{10}, y_3, \ldots, y_n)} \\ &\quad \cdots \frac{p(y_n|y_{10}, \ldots, y_{n-1,0})}{p(y_{n0}|y_{10}, \ldots, y_{n-1,0})} \cdot p(y_{10}, \ldots, y_{n0}) \,, \end{aligned} \tag{4.7}$$

an identity which is easily checked (Exercise 1). Here, $\mathbf{y}_0 = (y_{10}, \ldots, y_{n0})'$ is any fixed point in the support of $p(y_1, \ldots, y_n)$. Hence $p(y_1, \ldots, y_n)$ is determined by the full conditional distributions, since apart from the constant $p(y_{10}, \ldots, y_{n0})$ they are the only objects appearing on the right-hand side of (4.7). Hence the joint distribution is determined up to a proportionality constant. If $p(y_1, \ldots, y_n)$ is improper then this is, of course, the best we can do; if $p(y_1, \ldots, y_n)$ is proper then the fact that it integrates to 1 determines the constant. Perhaps most important is the constructive nature of (4.7): we can create $p(y_1, \ldots, y_n)$ simply by calculating the product of ratios. For more on this point, see Exercise 2.

When the number of areal units is very large (say, a regular grid of pixels associated with an image or a large number of small geographic regions), we do not seek to write down the joint distribution of the Y_i. Rather we prefer to work (and model) exclusively with the n corresponding full conditional distributions. In fact, from a spatial perspective, we would imagine that the full conditional distribution for Y_i would be more "local," that is, it should really depend only upon the neighbors of cell i. Adopting some definition of a neighbor structure (e.g., the one setting $W_{ij} = 1$ or 0 depending on whether i and j are adjacent or not), let ∂_i denote the set of neighbors of cell i.

Next suppose we specify a set of full conditional distributions for the Y_i such that

$$p(y_i|y_j, j \neq i) = p(y_i|y_j, j \in \partial_i) \tag{4.8}$$

A critical question to ask is whether a specification such as (4.8) uniquely determines a joint distribution for $Y_1, \ldots Y_n$. That is, we do not need to see the explicit form of this distribution. We merely want to be assured that if, for example, we implement a Gibbs sampler (see Subsection 5.3.1) to simulate realizations from the joint distribution, that there is indeed a unique stationary distribution for this sampler.

The notion of using *local* specification to determine a joint (or global) distribution in the form (4.8) is referred to as a *Markov random field* (MRF). There is by now a substantial literature in this area, with Besag (1974) being a good place to start. Geman and Geman (1984) provide the next critical step in the evolution, while Kaiser and Cressie (2000) offer a current view and provide further references. See also Rue and Held (2005) and references therein.

A critical definition in this regard is that of a *clique*. A clique is a set of cells (equivalently, indices) such that each element is a neighbor of every other element. With n cells, depending upon the definition of the neighbor structure, cliques can possibly be of size 1, 2, and so on up to size n. A *potential function* (or simply *potential*) of order k is a function of k arguments that is exchangeable in these arguments. The arguments of the potential would be the values taken by variables associated with the cells for a clique of size k. For continuous

Y_i, a customary potential on cliques of size $k = 2$ is $(Y_i - Y_j)^2$ when $i \sim j$. (We use the notation $i \sim j$ if i is a neighbor of j and j is a neighbor of i.) In fact, we may also view this potential as a sum of a potential on cliques of size $k = 1$, i.e., Y_i^2 with a potential on cliques of size $k = 2$, i.e., Y_iY_j. For, say, binary Y_i, a common potential on cliques of size $k = 2$ is

$$I(Y_i = Y_j) = Y_iY_j + (1 - Y_i)(1 - Y_j) ,$$

where again $i \sim j$ and I denotes the indicator function. Next, we define a *Gibbs distribution* as follows: $p(y_1, \ldots, y_n)$ is a Gibbs distribution if it is a function of the Y_i only through potentials on cliques. That is,

$$p(y_1, \ldots, y_n) \propto \exp\left\{\gamma \sum_k \sum_{\boldsymbol{\alpha} \in \mathcal{M}_k} \phi^{(k)}(y_{\alpha_1}, y_{\alpha_2}, \ldots, y_{\alpha_k})\right\} . \tag{4.9}$$

Here, $\phi^{(k)}$ is a potential of order k, $\mathcal{M}_k$ is the collection of all subsets of size k from $\{1, 2, \ldots, n\}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ indexes this set, and $\gamma > 0$ is a scale (or "temperature") parameter. If we only use cliques of size 1, we see that we obtain an independence model, evidently not of interest. When $k = 2$, we achieve spatial structure. In practice, cliques with $k = 3$ or more are rarely used, introducing complexity with little benefit. So, throughout this book, only cliques of order less than or equal to 2 are considered.

Informally, the *Hammersley-Clifford Theorem* (see Besag, 1974; also Clifford, 1990) demonstrates that if we have an MRF, i.e., if (4.8) defines a unique joint distribution, then this joint distribution is a Gibbs distribution. That is, it is of the form (4.9), with all of its "action" coming in the form of potentials on cliques. Cressie (1993, pp. 417–18) offers a proof of this theorem, and mentions that its importance for spatial modeling lies in its limiting the complexity of the conditional distributions required, i.e., full conditional distributions can be specified locally.

Geman and Geman (1984) provided essentially the converse of the Hammersley-Clifford Theorem. If we begin with (4.9) we have determined an MRF. As a result, they argued that to sample a Markov random field, one could sample from its associated Gibbs distribution, hence coining the term "Gibbs sampler."

For continuous data on $\Re^1$, a common choice for the joint distribution is a pairwise difference form

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2} \sum_{i,j} (y_i - y_j)^2 I(i \sim j)\right\} . \tag{4.10}$$

Distributions such as (4.10) will be the focus of the next section. For the moment, we merely note that it is a Gibbs distribution on potentials of order 1 and 2 and that

$$p(y_i \,\big|\, y_j, j \neq i) = N\left(\sum_{j \in \partial_i} y_i/m_i \,,\, \tau^2/m_i\right) , \tag{4.11}$$

where m_i is the number of neighbors of cell i. The distribution in (4.11) is clearly of the form (4.8) and shows that the mean of Y_i is the average of its neighbors, exactly the sort of local smoother we discussed in the section on spatial smoothers.

4.3 Conditionally autoregressive (CAR) models

Although they were introduced by Besag (1974) approximately 30 years ago, conditionally autoregressive (CAR) models have enjoyed a dramatic increase in usage only in the past decade or so. This resurgence arises from their convenient employment in the context of Gibbs sampling and more general Markov chain Monte Carlo (MCMC) methods for fitting certain classes of hierarchical spatial models (seen, e.g., in Section 6.4.3).

4.3.1 The Gaussian case

We begin with the Gaussian (or *autonormal*) case. Suppose we set

$$Y_i \mid y_j, j \neq i \sim N\left(\sum_j b_{ij} y_j \, , \, \tau_i^2\right), \; i = 1, \ldots, n \, . \tag{4.12}$$

These full conditionals are compatible, so through Brook's Lemma we can obtain

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2}\mathbf{y}' D^{-1}(I - B)\mathbf{y}\right\} \, , \tag{4.13}$$

where $B = \{b_{ij}\}$ and D is diagonal with $D_{ii} = \tau_i^2$. Expression (4.13) suggests a joint multivariate normal distribution for $\mathbf{Y}$ with mean $\mathbf{0}$ and variance matrix $\Sigma_{\mathbf{y}} = (I - B)^{-1}D$.

But we are getting ahead of ourselves. First, we need to ensure that $D^{-1}(I - B)$ is symmetric. The resulting conditions are

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \quad \text{for all} \;\; i, j \, . \tag{4.14}$$

Evidently, from (4.14), B need not be symmetric. Returning to our proximity matrix W (which we assume to be symmetric), suppose we set $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$. Then (4.14) is satisfied and (4.12) yields $p(y_i|y_j, j \neq i) = N\left(\sum_j w_{ij} y_j/w_{i+} \, , \, \tau^2/w_{i+}\right)$. Also, (4.13) becomes

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\mathbf{y}'(D_w - W)\mathbf{y}\right\} \, , \tag{4.15}$$

where D_w is diagonal with $(D_w)_{ii} = w_{i+}$.

Now a second aspect is noticed. $(D_w - W)\mathbf{1} = \mathbf{0}$, i.e., $\Sigma_{\mathbf{y}}^{-1}$ is singular, so that $\Sigma_{\mathbf{y}}$ does not exist and the distribution in (4.15) is improper. (The reader is encouraged to note the difference between the case of $\Sigma_{\mathbf{y}}^{-1}$ singular and the case of $\Sigma_{\mathbf{y}}$ singular. With the former we have a density function but one that is not integrable; effectively we have too many variables and we need a constraint on them to restore propriety. With the latter we have no density function but a proper distribution that resides in a lower dimensional space; effectively we have too *few* variables.) With a little algebra (4.15) can be rewritten as

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i \neq j} w_{ij}(y_i - y_j)^2\right\} \, . \tag{4.16}$$

This is a pairwise difference specification slightly more general than (4.10). But the impropriety of $p(\mathbf{y})$ is also evident from (4.16) since we can add any constant to all of the Y_i and (4.16) is unaffected; the Y_i are not "centered." A constraint such as $\sum_i Y_i = 0$ would provide the needed centering. Thus we have a more general illustration of a joint distribution that is improper, but has all full conditionals proper. The specification in (4.15) or (4.16) is often referred to as an *intrinsically autoregressive* (IAR) model.

As a result, $p(\mathbf{y})$ in (4.15) cannot be used as a model for data; data could not arise under an improper stochastic mechanism, and we cannot impose a constant center on randomly realized measurements. Hence, the use of an improper autonormal model must be relegated to a *prior* distributional specification. That is, it will be attached to random spatial effects introduced at the second stage of a hierarchical specification (again, see, e.g., Section 6.4.3).

The impropriety in (4.15) can be remedied in an obvious way. Redefine $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$ and choose ρ to make $\Sigma_{\mathbf{y}}^{-1}$ nonsingular. This is guaranteed if $\rho \in \left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$, where

$\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$; see Exercise 5; Moreover, since $tr(D_w^{-1/2} W D_w^{-1/2}) = 0 = \sum_{i=1}^n \lambda_{(i)}$, $\lambda_{(1)} < 0$, $\lambda_{(n)} > 0$, and 0 belongs to $\left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$.

Simpler bounds than those given above for the propriety parameter ρ may be obtained if we replace the adjacency matrix W by the scaled adjacency matrix $\widetilde{W} \equiv Diag(1/w_{i+})W$; recall $\widetilde{W}$ is not symmetric, but it will be row stochastic (i.e., all of its rows sum to 1). $\Sigma_{\mathbf{y}}^{-1}$ can then be written as $M^{-1}(I - \alpha\widetilde{W})$ where M is diagonal. Then if $|\alpha| < 1$, $I - \alpha\widetilde{W}$ is nonsingular. (See the SAR model of the next section, as well as Exercise 9;) Carlin and Banerjee (2003) show that $\Sigma_{\mathbf{y}}^{-1}$ is diagonally dominant and symmetric. But diagonally dominant symmetric matrices are positive definite (Harville, 1997), providing an alternative argument for the propriety of the joint distribution.

An elegant way to look at the propriety of CAR models is through the Gershgorin disk theorem (Golub and Van Loan, 2012; Horn and Johnson, 2012). This famous theorem of linear algebra focuses on so-called *diagonal dominance* and, in its simplest form, asserts that, for any symmetric matrix A, if all $a_{ii} > 0$ and $a_{ii} > \sum_{j \neq i} |a_{ij}|$, then A is positive definite. For instance, if $D_w^{-1}(I - B)$ is symmetric, then it is positive definite if, for each i, $\sum_{j \neq i} |B_{ij}| < 1$. With $D_w - \rho W$, a sufficient condition is that $|\rho| < 1$, weaker than the conditions created above. However, since we have positive definiteness for $\rho < 1$, impropriety at $\rho = 1$, this motivates us to examine the behavior of CAR models for ρ near 1, as we do below.

Returning to the unscaled situation, ρ can be viewed as an additional parameter in the CAR specification, enriching this class of spatial models. Furthermore, $\rho = 0$ has an immediate interpretation: the Y_i become independent $N(0, \tau^2/w_{i+})$. If ρ is not included, independence cannot emerge as a limit of (4.15). (Incidentally, this suggests a clarification of the role of τ^2, the variance parameter associated with the full conditional distributions: the magnitude of τ^2 should *not* be viewed as, in any way, *quantifying* the strength of spatial association. Indeed if all Y_i are multiplied by c, τ^2 becomes $c\tau^2$ but the strength of spatial association among the Y_i is clearly unaffected.) Lastly, $\rho \sum_j w_{ij} Y_j / w_{i+}$ can be viewed as a *reaction function*, i.e., ρ is the expected proportional "reaction" of Y_i to $\sum_j w_{ij} Y_j / w_{i+}$. (This interpretation is more common in the SAR literature (Section 4.4).)

With these advantages plus the fact that $p(\mathbf{y})$ (or the Bayesian posterior distribution, if the CAR specification is used to model constrained random effects) is now proper, is there any reason not to introduce the ρ parameter? In fact, the answer may be yes. Under $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$, the full conditional $p(y_i | y_j, j \neq i)$ becomes $N\left(\rho \sum_j w_{ij} y_j / w_{i+}\,,\, \tau^2 / w_{i+}\right)$. Hence we are modeling Y_i not to have mean that is an average of its neighbors, but some *proportion* of this average. Does this enable any sensible spatial interpretation for the CAR model? Moreover, does ρ calibrate very well with any familiar interpretation of "strength of spatial association"? Fixing $\tau^2 = 1$ without loss of generality, we can simulate CAR realizations for a given n, W, and ρ. We can also compute for these realizations a descriptive association measure such as Moran's I or Geary's C. Here we do not present explicit details of the range of simulations we have conducted. However, for a 10×10 grid using a first-order neighbor system, when $\rho = 0.8$, I is typically 0.1 to 0.15; when $\rho = 0.9$, I is typically 0.2 to 0.25; and even when $\rho = 0.99$, I is typically at most 0.5. It thus appears that ρ can mislead with regard to strength of association. Expressed in a different way, within a Bayesian framework, a prior on ρ that encourages a consequential amount of spatial association would place most of its mass near 1.

A related point is that if $p(\mathbf{y})$ is proper, the breadth of spatial pattern may be too limited. In the case where a CAR model is applied to random effects, an improper choice may actually enable wider scope for posterior spatial pattern. As a result, we do not take a position with regard to propriety or impropriety in employing CAR specifications (though

in the remainder of this text we do sometimes attempt to illuminate relative advantages and disadvantages).

Referring to (4.12), we may write the entire system of random variables as

$$\mathbf{Y} = B\mathbf{Y} + \boldsymbol{\epsilon}\,, \quad \text{or equivalently,} \tag{4.17}$$

$$(I - B)\mathbf{Y} = \boldsymbol{\epsilon}\,. \tag{4.18}$$

In particular, the distribution for $\mathbf{Y}$ induces a distribution for $\boldsymbol{\epsilon}$. If $p(\mathbf{y})$ is proper then $\mathbf{Y} \sim N(\mathbf{0}, (I-B)^{-1}D)$ whence $\boldsymbol{\epsilon} \sim N(\mathbf{0}, D(I-B)^T)$, i.e., the components of $\boldsymbol{\epsilon}$ are not independent. Also, $Cov(\boldsymbol{\epsilon}, \mathbf{Y}) = D$. The SAR specification in Section 4.4 reverses this specification, supplying a distribution for $\boldsymbol{\epsilon}$ which induces a distribution for $\mathbf{Y}$.

When $p(\mathbf{y})$ is proper we can appeal to standard multivariate normal distribution theory to interpret the entries in $\Sigma_{\mathbf{y}}^{-1}$. For example, $1/(\Sigma_{\mathbf{y}}^{-1})_{ii} = Var(Y_i|Y_j, j \neq i)$. Of course with $\Sigma_{\mathbf{y}}^{-1} = D^{-1}(I-B)$, $(\Sigma_{\mathbf{y}}^{-1})_{ii} = 1/\tau_i^2$ providing immediate agreement with (4.12). But also, if $(\Sigma_{\mathbf{y}}^{-1})_{ij} = 0$, then Y_i and Y_j are conditionally independent given $Y_k, k \neq i, j$, a fact you are asked to show in Exercise 10. Hence if any $b_{ij} = 0$, we have conditional independence for that pair of variables. Connecting b_{ij} to w_{ij} shows that the choice of neighbor structure implies an associated collection of conditional independences. With first-order neighbor structure, all we are asserting is a spatial illustration of the local Markov property (Whittaker, 1990, p. 68).

We conclude this subsection with four remarks. First, one can directly introduce a regression component into (4.12), e.g., a term of the form $\mathbf{x}_i'\boldsymbol{\beta}$. Conditional on $\boldsymbol{\beta}$, this does not affect the association structure that ensues from (4.12); it only revises the mean structure. However, we omit details here (the interested reader can consult Besag, 1974), since we will only use the autonormal CAR as a distribution for spatial random effects. These effects are added onto the regression structure for the mean on some transformed scale (again, see Section 6.4.3).

We also note that in suitable contexts it may be appropriate to think of $\mathbf{Y}_i$ as a vector of dependent areal unit measurements or, in the context of random effects, as a vector of dependent random effects associated with an areal unit. This leads to the specification of multivariate conditionally autoregressive (MCAR) models, which is the subject of Section 10.1. From a somewhat different perspective, $\mathbf{Y}_i$ might arise as $(Y_{i1}, \ldots, Y_{iT})^T$ where Y_{it} is the measurement associated with areal unit i at time t, $t = 1, \ldots, T$. Now we would of course think in terms of spatiotemporal modeling for Y_{it}. This is the subject of Section 11.7.

Thirdly, a (proper) CAR model can in principle be used for point-level data, taking w_{ij} to be, say, an inverse distance between points i and j. However, unlike the spatial prediction described in Section 2.4, now spatial prediction becomes *ad hoc*. That is, to predict at a new site Y_0, we might specify the distribution of Y_0 given $Y_1, \ldots, Y_n$ to be a normal distribution, such as a $N\left(\rho \sum_j w_{0j} y_j / w_{0+}\,,\ \tau^2/w_{0+}\right)$. Note that this determines the joint distribution of $Y_0, Y_1, \ldots, Y_n$. However, this joint distribution is *not* the CAR distribution that would arise by specifying the full conditionals for $Y_0, Y_1, \ldots, Y_n$ and using Brook's Lemma, as in constructing (4.15). In this regard, we cannot "marginalize" a CAR model. That is, suppose we specify a CAR model for, say, n areal units and we want a CAR model for a subset of them, say the first m. If we consider the multivariate normal distribution with upper left $m \times m$ block $(D^{-1}(I-B))_m^{-1}$, the inverse of this matrix need not look anything like the CAR model for these m units.

Finally, Gaussian Markov random fields can introduce proximities more general than those that we have discussed here. In particular, working with a regular lattice, there is much scope for further theoretical development. For instance, Rue and Held (2005, p. 114) describe the derivation of the following model weights based on the forward difference analogue of penalizing the derivatives of a surface used to specify the thin plate spline. They consider 12 neighbors of a given point. The north, east, south, and west neighbors

each receive a weight of +8, the northeast, southeast, southwest, and northwest neighbors, each receive a weight of −2 and the "two away" north, east, south, and west neighbors, each receive a weight of −1. Thus, the $w_{i+} = 20$. These weights would possibly viewed as unusual with regard to spatial smoothing, in particular the negative values, but, again, they do have a probabilistic justification through the two-dimensional random walk on the lattice. Moreover, they do play a role in Markov random field approximation to Gaussian processes. Some care needs to be taken with regard to edge specifications. See further discussion in Rue and Held (2005).

4.3.2 The non-Gaussian case

If one seeks to model the data directly using a CAR specification, then in many cases a normal distribution would not be appropriate. Binary response data and sparse count data are two examples. In fact, one might imagine any exponential family model as a first-stage distribution for the data. Here, we focus on the case where the Y_i are binary variables and present the so-called autologistic CAR model (historically, the Ising model; see Brush, 1967). This model has received attention in the literature; see, e.g., Heikkinen and Hogmander (1994), Hogmander and Møller (1995), and Hoeting et al. (2000). Ignoring covariates for the moment, as we did with the CAR models above, consider the joint distribution

$$\begin{aligned} p(y_1, y_2, ..., y_n; \psi) &\propto \exp(\psi \sum_{i,j} w_{ij} 1(y_i = y_j)) \\ &= \exp(\psi \sum_{i,j} w_{ij}(y_i y_j + (1 - y_i)(1 - y_j))). \end{aligned} \tag{4.19}$$

We immediately recognize this specification as a Gibbs distribution with a potential on cliques of order $k = 2$. Moreover, this distribution is always proper since it can take on only 2^n values. However, we will assume that ψ is an unknown parameter (how would we know it in practice?) and hence we will need to calculate the normalizing constant $c(\psi)$ in order to infer about ψ. But, computation of this constant requires summation over all of the 2^n possible values that $(Y_1, Y_2, ..., Y_n)$ can take on. Even for moderate sample sizes this will present computational challenges. Hoeting et al. (2000) propose approximations to the likelihood using a pseudo-likelihood and a normal approximation.

From (4.19) we can obtain the full conditional distributions for the Y_i's. In fact, $P(Y_i = 1|y_j, j \neq i) = e^{\psi S_{i,1}}/(e^{\psi S_{i,0}} + e^{\psi(S_{i,1})})$ where $S_{i,1} = \sum_{j\sim i} 1(y_j = 1)$ and $S_{i,0} = \sum_{j\sim i} 1(y_j = 0)$ and $P(Y_i = 0|y_j, j \neq i) = 1 - P(Y_i = 1|y_j, j \neq i)$. That is, $S_{i,1}$ is the number of neighbors of i that are equal to 1 and $S_{i,0}$ is the number of neighbors of i that are equal to 0. We can see the role that ψ plays; larger values of ψ place more weight on matching. This is most easily seen through $\log \frac{P(Y_i=1|y_j, j\neq i)}{P(Y_i=0|y_j, j\neq i)} = \psi(S_{i,1} - S_{i,0})$. Since the full conditional distributions take on only two values, there are no normalizing issues with them.

Bringing in covariates is natural on the log scale, i.e.,

$$\log \frac{P(Y_i = 1|y_j, j \neq i)}{P(Y_i = 0|y_j, j \neq i)} = \psi(S_{i,1} - S_{i,0}) + \mathbf{X}_i^T \boldsymbol{\beta}. \tag{4.20}$$

Solving for $P(Y_i = 1|y_j, j \neq i)$, we obtain

$$P(Y_i = 1 \mid y_j, j \neq i) = \frac{\exp\{\psi(S_{i,1} - S_{i,0}) + \mathbf{X}_i^T \boldsymbol{\beta}\}}{1 + \exp\{(S_{i,1} - S_{i,0}) + \mathbf{X}_i^T \boldsymbol{\beta}\}} .$$

Now, to update both ψ and $\boldsymbol{\beta}$, we will again need the normalizing constant, now $c(\psi, \boldsymbol{\beta})$. In fact, we leave as an exercise, the joint distribution of $(Y_1, Y_2, ..., Y_n)$ up to a constant.

The case where Y_i can take on one of several categorical values presents a natural extension to the autologistic model. If we label the (say) L possible outcomes as simply $1, 2, ..., L$, then we can define the joint distribution for $(Y_1, Y_2, ..., Y_n)$ exactly as in (4.19), i.e.,

$$p(y_1, y_2, ..., y_n; \psi) \propto \exp(\psi \sum_{i,j} w_{ij} 1(y_i = y_j)) \tag{4.21}$$

with w_{ij} as above. The distribution in (4.21) is referred to as a *Potts model* (Potts, 1952). Now the distribution takes on L^n values; now, calculation of the normalizing constant is even more difficult. Because of this challenge, fitting Potts models to data is rare in the literature; rather, it is customary to run a forward simulation using the Potts model since this only requires implementing a routine Gibbs sampler, updating the Y_i's (see Chapter 5) for a fixed ψ. However, one nice data analysis example is the allocation model in Green and Richardson (2002). There, the Potts model is employed as a random effects specification, in a disease mapping context (see Chapter 6), as an alternative to a CAR model.

4.4 Simultaneous autoregressive (SAR) models

Returning to (4.17), suppose that instead of letting $\mathbf{Y}$ induce a distribution for $\boldsymbol{\epsilon}$, we let $\boldsymbol{\epsilon}$ induce a distribution for $\mathbf{Y}$. Imitating usual autoregressive time series modeling, suppose we take the ϵ_i to be independent innovations. For a little added generality, assume that $\boldsymbol{\epsilon} \sim N\left(0, \tilde{D}\right)$ where $\tilde{D}$ is diagonal with $\left(\tilde{D}\right)_{ii} = \sigma_i^2$. (Note $\tilde{D}$ has no connection with D in Section 4.3; the B we use below may or may not be the same as the one we used in that section.) Analogous to (4.12), now $Y_i = \sum_j b_{ij} Y_j + \epsilon_i$, $i = 1, 2, ..., n$, with $\epsilon_i \sim N\left(0, \sigma_i^2\right)$ or, equivalently, $(I - B)\mathbf{Y} = \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ distributed as above. Therefore, if $(I - B)$ is full rank,

$$\mathbf{Y} \sim N\left(\mathbf{0}\,,\, (I - B)^{-1} \tilde{D}\, ((I - B)^{-1})^T\right)\,. \tag{4.22}$$

Also, $Cov(\boldsymbol{\epsilon}, \mathbf{Y}) = \tilde{D}(I - B)^{-1}$. If $\tilde{D} = \sigma^2 I$ then (4.22) simplifies to $\mathbf{Y} \sim N\left(\mathbf{0}\,,\, \sigma^2 \left[(I - B)(I - B)^T\right]^{-1}\right)$. In order that (4.22) be proper, $I - B$ must be full rank but not necessarily symmetric. Two choices are most frequently discussed in the literature (e.g., Griffith, 1988). The first assumes $B = \rho W$, where W is a so-called contiguity matrix, i.e., W has entries that are 1 or 0 according to whether or not unit i and unit j are direct neighbors (with $w_{ii} = 0$). So W is our familiar first-order neighbor proximity matrix. Here ρ is called a *spatial autoregression parameter* and, evidently, $Y_i = \rho \sum_j Y_j I(j \in \partial_i) + \epsilon_i$, where ∂_i denotes the set of neighbors of i. In fact, any symmetric proximity matrix can be used and, paralleling the discussion below (4.15), $I - \rho W$ will be nonsingular if $\rho \in \left(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$ where now $\lambda_{(1)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of W. As a weaker conclusion, if W is symmetric, we can apply the diagonal dominance result from Section 4.3.1. Now, if $\rho \sum_{j \neq i} w_{ij} < 1$ for each i, i.e., $\rho < \min \frac{1}{w_{i+}}$, we have positive definiteness, hence nonsingularity.

Alternatively, W can be replaced by $\widetilde{W}$ where now, for each i, the ith row has been normalized to sum to 1. That is, $\left(\tilde{W}\right)_{ij} = w_{ij}/w_{i+}$. Again, $\widetilde{W}$ is not symmetric, but it is row stochastic, i.e., $\widetilde{W}\mathbf{1} = \mathbf{1}$. If we set $B = \alpha\widetilde{W}$, α is called a *spatial autocorrelation parameter* and, were W a contiguity matrix, now $Y_i = \alpha \sum_j Y_i I(j \in \partial_i)/w_{i+} + \epsilon_i$. With a very regular grid the w_{i+} will all be essentially the same and thus α will be a multiple of ρ. But, perhaps more importantly, with $\widetilde{W}$ row stochastic the eigenvalues of $\widetilde{W}$ are all less than or equal to 1 (i.e., $\max |\lambda_i| = 1$). Thus $I - \alpha\widetilde{W}$ will be nonsingular if $\alpha \in (-1, 1)$, justifying referring to α as an autocorrelation parameter; see Exercise 9.

A SAR model is customarily introduced in a regression context, i.e., the *residuals* $\mathbf{U} = \mathbf{Y} - X\boldsymbol{\beta}$ are assumed to follow a SAR model, rather than $\mathbf{Y}$ itself. But then, following (4.17), if $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, we obtain the attractive form

$$\mathbf{Y} = B\mathbf{Y} + (I - B)\, X\boldsymbol{\beta} + \boldsymbol{\epsilon}\,. \tag{4.23}$$

Expression (4.23) shows that $\mathbf{Y}$ is modeled through a component that provides a spatial weighting of neighbors and a component that is a usual linear regression. If B is the zero matrix we obtain an OLS regression; if $B = I$ we obtain a purely spatial model.

We note that from (4.23) the SAR model does not introduce any spatial effects; the errors in (4.23) are independent. Expressed in a different way, if we modeled $\mathbf{Y} - X\boldsymbol{\beta}$ as $\mathbf{U} + \mathbf{e}$ with $\mathbf{e}$ independent errors, we would have $\mathbf{U} + \mathbf{e} = B\mathbf{U} + \boldsymbol{\epsilon} + \mathbf{e}$ and $\boldsymbol{\epsilon} + \mathbf{e}$ would result in a redundancy. Equivalently, if we write $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, we see, from the distribution of $B\mathbf{U}$, that both terms on the right side are driven by the same variance component, σ^2_ϵ. As a result, in practice a SAR specification is not used in conjunction with a GLM. To introduce $\mathbf{U}$ as a vector of spatial adjustments to the mean vector, a transformed scale creates redundancy between the independent Gaussian error in the definition of the U_i and the stochastic mechanism associated with the conditionally independent Y_i.

We briefly note the somewhat related spatial modeling approach of Langford et al. (1999). Rather than modeling the residual vector $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, they propose that $\mathbf{U} = \tilde{B}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2 I\right)$, i.e., that $\mathbf{U}$ be modeled as a spatially motivated linear combination of independent variables. This induces $\Sigma_U = \sigma^2 \tilde{B}\tilde{B}^T$. Thus, the U_i and hence the Y_i will be dependent and given $\tilde{B}$, $cov\left(Y_i, Y_{i'}\right) = \sigma^2 \sum_j \tilde{b}_{ij}\tilde{b}_{i'j}$. If $\tilde{B}$ arises through some proximity matrix W, the more similar rows i and i' of W are, the stronger the association between Y_i and $Y_{i'}$. However, the difference in nature between this specification and that in (4.23) is evident. To align the two, we would set $(I - B)^{-1} = \tilde{B}$, i.e. $B = I - \tilde{B}^{-1}$ (assuming $\tilde{B}$ is of full rank). $I - \tilde{B}^{-1}$ would not appear to have any interpretation through a proximity matrix.

Perhaps the most important point to note with respect to SAR models is that they are well suited to maximum likelihood estimation but not at all for MCMC fitting of Bayesian models. That is, the log likelihood associated with (4.23) (assuming $\tilde{D} = \sigma^2 I$) is

$$\frac{1}{2}\log\left|\sigma^{-1}\left(I - B\right)\right| - \frac{1}{2\sigma^2}\left(\mathbf{Y} - X\boldsymbol{\beta}\right)^T \left(I - B\right)\left(I - B\right)^T \left(\mathbf{Y} - X\boldsymbol{\beta}\right)\,. \tag{4.24}$$

Though B will introduce a regression or autocorrelation parameter, the quadratic form in (4.24) is quick to calculate (requiring no inverse) and the determinant can usually be calculated rapidly using diagonally dominant, sparse matrix approximations (see, e.g., Pace and Barry, 1997a,b). Thus maximization of (4.24) can be done iteratively but, in general, efficiently.

Also, note that while the form in (4.24) can certainly be extended to a full Bayesian model through appropriate prior specifications, the absence of a hierarchical form with random effects implies straightforward Bayesian model fitting as well. Indeed, the general spatial slice Gibbs sampler (see Appendix Section A.2, or Agarwal and Gelfand, 2002) can easily handle this model. However, suppose we attempt to introduce SAR random effects in some fashion. Unlike CAR random effects that are defined through full conditional distributions, the full conditional distributions for the SAR effects have no convenient form. For large n, computation of such distributions using a form such as (4.22) will be expensive.

SAR models as in (4.23) are frequently employed in the spatial econometrics literature. With point-referenced data, B is taken to be ρW where W is the matrix of inter-point distances. Likelihood-based inference can be implemented in the `spdep` package in `R` as well as more specialized software, such as that from the Spatial Analysis Laboratory (`sal.agecon.uiuc.edu`). Software for large data sets is supplied there, as well as through the website of Prof. Kelley Pace, `www.spatial-statistics.com`. An illustrative example is provided in Exercise 12.

4.4.1 CAR versus SAR models

Cressie (1993, pp. 408–10) credits Brook (1964) with being the first to make a distinction between the CAR and SAR models, and offers a comparison of the two. To begin with, we may note from (4.13) and (4.22) that, under propriety, the two forms are equivalent if and only if

$$(I - B)^{-1}D = (I - \tilde{B})^{-1}\tilde{D}((I - \tilde{B})^{-1})^T ,$$

where we use the tilde to indicate matrices in the SAR model. Cressie then shows that any SAR model can be represented as a CAR model (since D is diagonal, we can straightforwardly solve for B), but gives a counterexample to prove that the converse is not true. Since all SAR models are proper while we routinely employ improper CAR models, it is not surprising that the latter is a larger class.

For the "proper" CAR and SAR models that include spatial correlation parameters ρ, Wall (2004) shows that the correlations between neighboring regions implied by these two models can be rather different; in particular, the first-order neighbor correlations increase at a slower rate as a function of ρ in the CAR model than they do for the SAR model. (As an aside, she notes that these correlations are not even monotone for $\rho < 0$, another reason to avoid negative spatial correlation parameters.) Also, correlations among pairs can switch in nonintuitive ways. For example, when working with the adjacency relationships generated by the lower 48 contiguous U.S. states, she finds that when $\rho = .49$ in the CAR model, $Corr(Alabama, Florida) = .20$ and $Corr(Alabama, Georgia) = .16$. But when ρ increases to .975, we instead get $Corr(Alabama, Florida) = .65$ and $Corr(Alabama, Georgia) = .67$, a slight reversal in ordering.

4.4.2 STAR models

In the literature SAR models have frequently been extended to handle spatiotemporal data. The idea is that in working with proximity matrices, we can define neighbors in time as well as in space. Figure 4.4 shows a simple illustration with 9 areal units, 3 temporal units for each areal unit yielding $i = 1, \ldots, 9$, $t = 1, 2, 3$, labeled as indicated.

The measurements Y_{it} are spatially associated at each fixed t. But also, we might seek to associate, say, Y_{i2} with Y_{i1} and Y_{i3}. Suppose we write Y as the 27×1 vector with the first nine entries at $t = 1$, the second nine at $t = 2$, and the last nine at $t = 3$. Also let $W_S = BlockDiag(W_1, W_1, W_1)$, where

$$W_1 = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} .$$

Then W_S provides a spatial contiguity matrix for the Y's. Similarly, let $W_T = \begin{pmatrix} 0 & W_2 & 0 \\ W_2 & 0 & W_2 \\ 0 & W_2 & 0 \end{pmatrix}$, where $W_2 = I_{9\times 9}$. Then W_T provides a *temporal* contiguity matrix for the Y's. But then, in our SAR model we can define $B = \rho_s W_S + \rho_t W_T$. In fact, we can

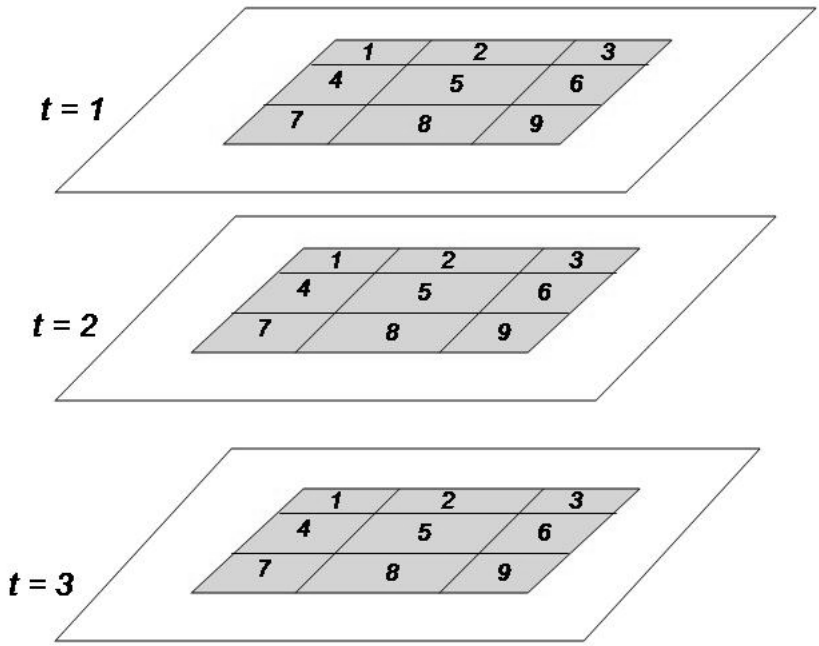

Figure 4.4 *Illustration of spatiotemporal areal unit setting for STAR model.*

also introduce $\rho_{ST} W_S W_T$ into B and note that

$$W_S W_T = \begin{pmatrix} 0 & W_1 & 0 \\ W_1 & 0 & W_1 \\ 0 & W_1 & 0 \end{pmatrix}.$$

In this way, we introduce association across both space and time. For instance Y_{21} and Y_{41} affect the mean of Y_{12} (as well as affecting Y_{11}) from W_S by itself. Many more possibilities exist. Models formulated through such more general definitions of B are referred to as *spatiotemporal autoregressive* (STAR) models. See Pace et al. (2000) for a full discussion and development. The interpretation of the ρ's in the above example measures the relative importance of first-order spatial neighbors, first-order temporal neighbors, and first-order spatiotemporal neighbors.

4.5 Computer tutorials

In this section we outline the use of some GIS functions in R for obtaining neighborhood (adjacency) matrices, computing Moran's and Geary's statistic and fitting CAR and SAR models using traditional maximum likelihood techniques, and mapping the results for certain classes of problems. Here we confine ourselves to the modeling of Gaussian data on areal units. For the following illustrations we will load the following libraries and use functions therein:

```
> library(maps)
> library(maptools)
> library(spdep)
> library(classInt) ## Will be used for plotting maps later
> library(RColorBrewer) ## Will be used for colored maps
```

4.5.1 Adjacency matrices from maps using `spdep`

Consider, for example, the SAT exam scores data from the lower 48 contiguous states of the U.S. and the District of Columbia. We can construct this map using the `maps` and `maptools` packages. The idea is to create a sequence of data structures that will eventually produce an adjacency matrix. We execute the following commands

```
> usa.state = map(database="state", fill=TRUE, plot=FALSE)
> state.ID <- sapply(strsplit(usa.state$names, ":"), function(x) x[1])
> usa.poly = map2SpatialPolygons(usa.state, IDs=state.ID)
> usa.nb = poly2nb(usa.poly)
> usa.adj.mat = nb2mat(usa.nb, style="B")
```

The object `"usa.state"` is returned by the `"map()"` function. Next, we extract the state ID's from the `"state"` database and use them to construct a `"SpatialPolygon"` object called `"usa.poly"`, which is then coverted to a neighborhood object `"usa.nb"` using the `spdep` function `poly2nb`. Finally the `nb2mat` function (also in `spdep`) produces the adjacency matrix. The option `style="B"` produces the basic binary coding. Therefore, `usa.adj.mat` produced above is a 49×49 matrix whose (i, j)-th entry is equal to 1 if i is a neighbor of j and 0 otherwise. All diagonal entries are 0. The option `style="W"` produces a row-normalized adjacency matrix. Other options are available and left to the reader to explore.

For constructing adjacency matrices for counties in a U.S. state, the above code needs to be modified slightly because of the way the `"state"` and `"county"` databases list their entries. In the `"state"` database, the states are listed by simply their names (e.g. `"maryland"`) when they make up a single region or are separated by a `":"` in cases when they are split into subregions (e.g. `"michigan:north"` and `"michigan:south"`). This can be easily checked by typing `"usa.state$names"` after the `"usa.state"` object has been created as above. Counties within a state, on the other hand, are listed by the name of the state followed by the name of the county (e.g. `"minnesota,hennepin"`). Therefore, to produce the county neighborhood matrix for the State of Minnesota, we execute

```
> mn.county = map("county", "minnesota", fill=TRUE, plot=FALSE)
> county.ID <- sapply(strsplit(mn.county$names, ","), function(x) x[2])
> mn.poly = map2SpatialPolygons(mn.county, IDs=county.ID)
> mn.nb = poly2nb(mn.poly)
> mn.adj.mat = nb2mat(mn.nb, style="B")
```

Note the different specification in the way the `"strsplit()"` function is implemented for getting the county identifiers. The rest of the code is fairly similar to that for the state adjacencies. Neighbors of any given county can be easily found from the adjacency matrix. For example, the neighbors of Winona county in Minnesota can be found as

```
> mn.region.id <- attr(mn.nb, "region.id")
> winona.neighbors.index = mn.nb[[match("winona", mn.region.id)]]
> winona.neighbors = rownames(mn.adj.mat[winona.neighbors.index,])
> winona.neighbors
[1] "fillmore" "houston"  "olmsted"  "wabasha"
```

Note: Since the region is restricted to Minnesota, this lists Winona's adjacent counties in Minnesota only. Winona has three other adjacent counties in Wisconsin: Buffalo, Trempealeau and La Crosse.

One could also create adjacency matrices from external shapefiles by executing

```
> mn.map.shp = readShapeSpatial("minnesota.shp")
> mn.nb.shp = poly2nb(mn.map.shp)
> mn.adj.mat.shp = nb2mat(mn.nb, style="B")
```

However, the adjacency matrices obtained from external shapefiles need not be identical to those obtained from R's map databases. In fact, the rows and columns of the neighborhood matrix `mn.adj.mat.shp` obtained from the ESRI shapefiles for Minnesota will not correspond to `mn.adj.mat` from R's maps. Nevertheless, these can be easily brought to the same order:

```
> ordered.index =  order(as.character(mn.map$NAME))
> mn.adj.mat.shp = mn.adj.mat.shp[ordered.index, ordered.index]
```

These are now in the same order. However, they are still not quite identical with 99.87% of the entries in agreement. This happens because the polygons in the ESRI shapefiles slightly differ from those in the `maps` package in R.

4.5.2 Moran's I and Geary's C in `spdep`

We can subsequently invoke the `moran.test` and `geary.test` functions in the `spdep` package to obtain Moran's I and Geary's C statistics. Let us illustrate using the SAT scores data presented in Figure 4.1. We first use the `nb` object `usa.nb` and convert it to a `listw` object in R.

```
> usa.listw = nb2listw(usa.nb, style="W")
```

The option `style="W"` takes a 0-1 neighbors list, where regions are either listed as neighbors or are absent, and creates row-normalized weights. We read the SAT scores using the dataset available in `http://www.biostat.umn.edu/~brad/data2.html`.

```
state.sat.scores = read.table("state-sat.dat", header=T)
```

Next, we use the `moran.test` function to obtain

```
> moran.test(state.sat.scores$VERBAL, listw=usa.listw),
```

which gives us the calculated Moran's I and the associated variance of the estimate. This yields a sample estimate of 0.6125 with an associated standard error of 0.0979. Geary's C can be computed analogously using the `geary.test` function:

```
geary.test(state.sat.scores$VERBAL, listw=usa.listw),
```

which yields the sample estimate and standard error of 0.3577 and 0.0984, respectively. For maps with "islands" or regions without neighbors, we need to set `zero.policy=TRUE` in the `nb2listw()` function. This ensures that weight vectors of zero length are applied to islands, i.e., regions without any neighbor in the neighbors list. Otherwise, the program terminates with an error.

4.5.3 SAR and CAR model fitting using `spdep` *in* R

We now turn to fitting spatial autoregression models using available functions in the `spdep` package. A convenient illustration is offered by the SIDS (sudden infant death syndrome) data, analyzed by Cressie and Read (1985), Cressie and Chan (1989), Cressie (1993, Sec. 6.2) and Kaluzny et al. (1998, Sec. 5.3), and already loaded into the `spdep` package. This dataset contains counts of SIDS deaths from 1974 to 1978 and counts from 1979 to 1983 along with related covariate information for the 100 counties in the U.S. State of North Carolina.

The dataset can be read from a shapefile `sids.shp` and an `nb` object can be constructed from a GAL file `ncCC89` containing the data analyzed by Cressie and Chan (1989). This data relates to counts aggregated from 1979 to 1983. Both these files are included in `spdep`. We execute the following steps:

```
> nc.sids <- readShapePoly(system.file("etc/shapes/sids.shp", package=
  "spdep")[1],
```

```
+       ID="FIPSNO", proj4string=CRS("+proj=longlat +ellps=clrk66"))
> rn <- sapply(slot(nc.sids, "polygons"), function(x) slot(x, "ID"))
> ncCC89.nb <- read.gal(system.file("etc/weights/ncCC89.gal", package=
  "spdep")[1],
+ region.id=rn)
```

The first step produces a `SpatialPolygonsDataFrame` object `nc.sids`, while the second step produces the region IDs and stores them in `rn`. The third step uses these region IDs to produce an `nb` object by directly reading from the GAL file. We next use a Freeman-Tukey transformation to produce the transformed rates and append them to the `nc.sids` object.

```
> nc.sids.rates.FT = sqrt(1000)*(sqrt(nc.sids$SID79/nc.sids$BIR79)
+                       + sqrt((nc.sids$SID79 + 1)/nc.sids$BIR79))
> nc.sids$rates.FT = nc.sids.rates.FT
```

We wish to regress these rates on the non-white birth rates over the same period. This variable is available as `NWBIR79` in the `nc.sids` object. We will use the Freeman-Tukey transformed birth rates:

```
> nc.sids.nwbir.FT = sqrt(1000)*(sqrt(nc.sids$NWBIR79/nc.sids$BIR79)
+                       + sqrt((nc.sids$NWBIR79 + 1)/nc.sids$BIR79))
> nc.sids$nwbir.FT = nc.sids.nwbir.FT
```

Maximum likelihood estimation of (4.23), which has the likelihood in (4.24), can be carried out using the `errorsarlm()` or, equivalently, the `spautolm()` function in `spdep`. These functions produce the same output. Below we demonstrate the latter. We first create a `listw` object using the 0-1 adjacency structure

```
> ncCC89.listw = nb2listw(ncCC89.nb, style="B", zero.policy=TRUE)
```

Note that the `zero.policy=TRUE` is required here because the county shapefile in `spdep` lists two counties, Dare and Hyde, as having zero neighbors. These counties are situated in coastal North Carolina and are adjacent to substantial bodies of water. These two counties can be identified as

```
> nc.county.id = attr(ncCC89.nb, "region.id")
> nc.no.neighbors = card(ncCC89.nb)
> nc.islands = as.character(nc.sids[card(ncCC89.nb) == 0, ]$NAME)
> nc.islands
[1] "Dare" "Hyde"
```

We now estimate the SAR model in (4.23) using

```
> nc.sids.sar.out = spautolm(rates.FT~ nwbir.FT, data=nc.sids, family="SAR",
+                            listw=ncCC89.listw, zero.policy=TRUE)
> summary(nc.sids.sar.out)
Call:
spautolm(formula = rates.FT ~ nwbir.FT, data = nc.sids, listw = ncCC89.listw,
    family = "SAR", zero.policy = TRUE)

Residuals:
       Min         1Q     Median         3Q        Max
-2.2660489 -0.4281394  0.0043310  0.4978178  2.5164979

Regions with no neighbours included:
 37055 37095

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.5216462  0.2570334  9.8106  < 2e-16
nwbir.FT    0.0124001  0.0071867  1.7254  0.08445
```

```
Lambda: 0.05206 LR test value: 2.218 p-value: 0.13641

Log likelihood: -123.6898
ML residual variance (sigma squared): 0.68697, (sigma: 0.82883)
Number of observations: 100
Number of parameters estimated: 4
AIC: 255.38
```

A CAR model can also be estimated by setting `family=CAR` in the above. We see the presence of a significantly positive intercept. The regression coefficient corresponding to the non-white birth rates seems to indicate a positive correlation but is not statistically significant. These results seem to be consistent with the estimates from ordinary least squares. We point out that an analysis of the 1974–1978 data by Kaluzny et al. (1998, Sec. 5.3) using the `S+ Spatial Stats` software, which was presented in the earlier edition of this book, returned a significant coefficient for birth rates. However, the dataset used there was different, the adjacency matrix of the state was modified to split a county into three regions and the weights used for the adjacency matrix also were different. The spatial autocorrelation parameter, denoted by `Lambda` in the output summary, is also not significant.

Once the estimates from the model have been obtained, we may wish to plot the fitted rates and compare them with the raw data. We first compute the fitted rates and append them to the `nc.sids` object. We then plot the fitted rates on a map.

```
> nc.sids$fitted.sar = fitted(nc.sids.sar.out)
> brks = c(0, 2.0, 3.0, 3.5, 6.0)
> color.pallete = rev(brewer.pal(4,"RdBu"))
> class.fitted = classIntervals(var=nc.sids$fitted.sar, n=4, style="fixed",
+                               fixedBreaks=brks, dataPrecision=4)
> color.code.fitted = findColours(class.fitted, color.pallete)
> plot(nc.sids, col=color.code.fitted)
```

The raw rates available as `nc.sids$rates.FT` can be plotted analogously. The resulting maps, with an added legend, are displayed in Figure 4.5. Although the spatial autocorrelation in the data was found to be modest, the fitted values from the SAR model clearly show the smoothing. Both the maps have the same color scheme.

Instead of defining a neighborhood structure completely in terms of spatial adjacency on the map, we may want to construct neighbors using a distance function. For example, given centroids of the various regions, we could identify regions as neighbors if and only if their intercentroidal distance is below a particular threshold. We illustrate using a dataset offering neighborhood-level information on crime, mean home value, mean income, and other variables for 49 neighborhoods in Columbus, OH, during 1980. More information on these data is available from Anselin (1988, p.189), or in Exercise 12. The data can be downloaded from `www.biostat.umn.edu/~brad/data/Columbus.dat` but that is not needed as it is available within the `spdep` package.

We begin by reading a Columbus shapefile and creating a `"SpatialPolygonsDataFrame"` object.

```
> library(spdep)
> columbus.poly <- readShapePoly(system.file("etc/shapes/columbus.shp",
+                                package="spdep")[1])
```

Suppose we would like to have regions with intercentroidal distances less than half the maximum intercentroidal distance as neighbors. We first construct an object, `columbus.coords`, that contain the centroids of the different regions.

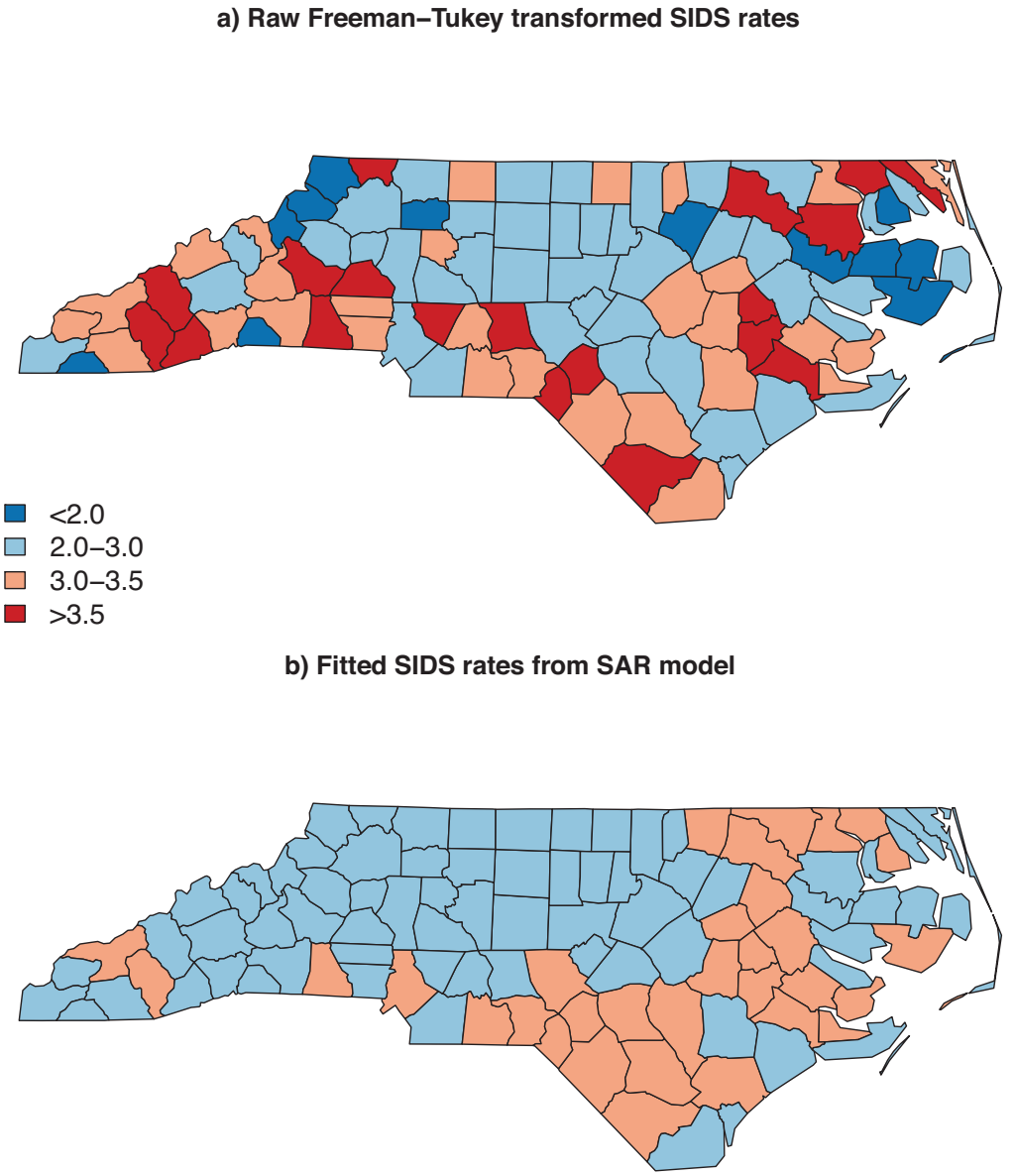

Figure 4.5 *Unsmoothed raw (a) and spatially smoothed fitted (b) rates, North Carolina SIDS data.*

```
> columbus.coords = coordinates(columbus.poly)
```

A particularly useful function provided by `spdep` is the `dnearneigh()` function. This function can take a matrix of coordinates along with specified lower and upper distance bounds as inputs and returns a list of neighbors categorized by all points that are within the specified distance threshold from each other. Note that the function uses Euclidean distances, which means that the coordinates must be projected onto a plane (and cannot be in terms of latitude and longitude) before the function is applied. There is, however, a potential pitfall of directly using the `dnearneigh()` function in that it may generate "islands" (i.e., regions with no neighbors) unless we are judicious about our choice of the upper distance bound (the lower bound is usually set to zero). One way to circumvent this problem is to first apply

the k-nearest neighbors algorithm to the coordinates and then to create an neighborhood list from these k-nearest neighbors:

```
> columbus.knn = knearneigh(columbus.coords)
> columbus.knn2nb = knn2nb(columbus.knn)
```

Next, the Euclidean distances between the neighbors are constructed by applying the `nbdists()` function to the k-nearest neighbor list. This returns the nearest-neighbor distances.

```
> columbus.dist.list = nbdists(columbus.knn2nb, columbus.coords)
> columbus.dist.vec = unlist(columbus.dist.list)
```

where the second step converts the list data structure into a vector. Next, we find the maximum of the nearest neighbor distances and pass it to the `dnearneigh()` function as an input for the upper bound.

```
> columbus.dist.max = max(columbus.dist.vec)
> columbus.dnn.nb = dnearneigh(columbus.coords, 0, columbus.dist.max)
```

This ensures that there are no islands. Next, we create a `listw` object from the returned `dnearneigh()` function and estimate the SAR model.

```
> columbus.dnn.listw = nb2listw(columbus.dnn.nb, style="B", zero.policy=TRUE)
> columbus.dnn.sar.out = spautolm(CRIME~HOVAL+INC, data=columbus.poly,
+                   family="SAR", listw=columbus.dnn.listw, zero.policy=TRUE)
> summary(columbus.dnn.sar.out)

Call: spautolm(formula = CRIME ~ HOVAL + INC, data = columbus.poly,
    listw = columbus.dnn.listw, family = "SAR", zero.policy = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-31.18714  -4.18864  -0.24961   6.10122  22.69041

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 50.229861   5.422987  9.2624 < 2.2e-16
HOVAL       -0.251366   0.079881 -3.1468  0.001651
INC         -0.826528   0.298414 -2.7697  0.005610

Lambda: 0.11302 LR test value: 19.536 p-value: 9.8716e-06

Log likelihood: -177.6092
ML residual variance (sigma squared): 73.306, (sigma: 8.5619)
Number of observations: 49
Number of parameters estimated: 5
AIC: 365.22
```

We see that both house value (HOVAL) and income (INC) have a significant negative impact on crime, which is not unexpected. The spatial autocorrelation parameter `Lambda` is also significant, which indicates strong spatial dependence in the data. Note, however, that this significance of the spatial dependence may be an artefact of an undesirably dense connectedness structure imposed by setting the upper distance bound to the maximum intercentroidal distance. To mitigate this effect one can set some proportion of the maximum distance as the distance upper bound. For example, we can set the upper bound to be `0.25*columbus.dist.max` in the above code and repeat the subsequent steps. This specification generates islands and `zero.policy=TRUE` option is required to obtain estimates from the SAR model. We do not show this output but it is similar to the above. While

the regression coefficients do not change substantially and are still significant, the spatial autocorrelation parameter is no longer statistically significant (p-value is approximately 0.25257). The result from the CAR model (obtained by setting `family="CAR"`) is also very similar.

4.6 Exercises

1. Verify Brook's Lemma, equation (4.7).
2.(a) To appreciate how Brook's Lemma works, suppose Y_1 and Y_2 are both binary variables, and that their joint distribution is defined through conditional logit models. That is,

$$\log \frac{P(Y_1 = 1|Y_2)}{P(Y_1 = 0|Y_2)} = \alpha_0 + \alpha_1 Y_2 \quad \text{and} \quad \log \frac{P(Y_2 = 1|Y_1)}{P(Y_2 = 0|Y_1)} = \beta_0 + \beta_1 Y_1 \ .$$

   Obtain the joint distribution of Y_1 and Y_2.

   (b) This result can be straightforwardly extended to the case of more than two variables, but the details become increasingly clumsy. Illustrate this issue in the case of *three* binary variables, Y_1, Y_2, and Y_3.
3. Returning to (4.13) and (4.14), let $B = ((b_{ij}))$ be an $n \times n$ matrix with positive elements; that is, $b_{ij} > 0$, $\sum_j b_{ij} \leq 1$ for all i, and $\sum_j b_{ij} < 1$ for at least one i. Let $D = Diag\left(\tau_i^2\right)$ be a diagonal matrix with positive elements τ_i^2 such that $D^{-1}(I - B)$ is symmetric; that is, $b_{ij}/\tau_i^2 = b_{ji}/\tau_j^2$, for all i, j. Show that $D^{-1}(I - B)$ is positive definite.
4. Looking again at (4.13), obtain a simple sufficient condition on B such that the CAR specification with precision matrix $D^{-1}(I - B)$ is a pairwise difference specification, as in (4.16).
5. Show that, for W symmetric, $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$ is positive definite (thus resolving the impropriety in (4.15)) if $\rho \in \left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$, where $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$.
6. Show that if all entries in W are nonnegative and $D_w - \rho W$ is positive definite with $0 < \rho < 1$, then all entries in $(D_w - \rho W)^{-1}$ are nonnegative.
7. Under a proper CAR model for $\mathbf{Y}$, i.e., with $\Sigma_{\mathbf{y}} = D_w - \rho W$, obtain the correlation and covariance between Y_i and Y_j.
8. Obtain the joint distribution, up to normalizing constant, for $(Y_1, Y_2, ..., Y_n)$ under (4.20). Hint: You might try to guess it but Brook's Lemma can be used as well.
9. Recalling the SAR formulation using the scaled adjacency matrix $\widetilde{W}$ just below (4.22), prove that $I - \alpha\widetilde{W}$ will be nonsingular if $\alpha \in (-1, 1)$, so that α may be sensibly referred to as an "autocorrelation parameter."
10. In the setting of Subsection 4.3.1, if $(\Sigma_{\mathbf{y}}^{-1})_{ij} = 0$, then show that Y_i and Y_j are conditionally independent given $Y_k, k \neq i, j$.
11. The file `www.biostat.umn.edu/~brad/data/state-sat.dat` gives the 1999 state average SAT data (part of which is mapped in Figure 4.1).

    (a) Use the `spautolm` function to fit the SAR model of Section 4.4, taking the verbal SAT score as the response Y and the percent of eligible students taking the exam in each state as the covariate X. Do this analysis twice: first using binary weights and then using row-normalized weights. Is the analysis sensitive to these choices of weights? Is knowing X helpful in explaining Y?

    (b) Using the `maps` library in `R`, draw choropleth maps similar to Figure 4.1 of both the fitted verbal SAT scores and the spatial residuals from the SAR model. Is there

evidence of spatial correlation in the response Y once the covariate X is accounted for?

(c) Repeat your SAR model analysis above, again using `spautolm` but now assuming the CAR model of Section 4.3. Compare your estimates with those from the SAR model and interpret any changes.

(d) One might imagine that the percentage of eligible students taking the exam should perhaps affect the variance of our model, not just the mean structure. To check this, refit the SAR model replacing your row-normalized weights with weights equal to the reciprocal of the percentage of students taking the SAT. Is this model sensible?

12. Consider the data `www.biostat.umn.edu/~brad/data/Columbus.dat`, taken from Anselin (1988, p. 189) and also available within the `spdep R` package (but with possibly different variable names). These data record crime information for 49 neighborhoods in Columbus, OH, during 1980. Variables measured include NEIG, the neighborhood id value (1–49); HOVAL, its mean housing value (in \$1,000); INC, its mean household income (in \$1,000); CRIME, its number of residential burglaries and vehicle thefts per thousand households; OPEN, a measure of the neighborhood's open space; PLUMB, the percentage of housing units without plumbing; DISCBD, the neighborhood centroid's distance from the central business district; X, an x-coordinate for the neighborhood centroid (in arbitrary digitizing units, not polygon coordinates); Y, the same as X for the y-coordinate; AREA, the neighborhood's area; and PERIM, the perimeter of the polygon describing the neighborhood.

(a) Use `spdep` in `R` to construct adjacency matrices for the neighborhoods of Columbus based upon centroid distances less than

i. 25% of the maximum intercentroidal distances;
ii. 50% of the maximum intercentroidal distances;
iii. 75% of the maximum intercentroidal distances.

(b) For each of the three spatial neighborhoods constructed above, use the `spautolm` function to fit SAR models with CRIME as the dependent variable, and HOVAL, INC, OPEN, PLUMB, and DISCBD as the covariates. Compare your results and interpret your parameter estimates in each case.

(c) Repeat your analysis by setting $B = \rho W$ in equation (4.23) with W_{ij} the Euclidean distance between location i and location j.

(d) Repeat part (b) for CAR models. Compare your estimates with those from the SAR model and interpret them.