CHAPTER 4

# MULTILEVEL STATISTICAL MODELS AND ECOLOGICAL SCALING

RICHARD A. BERK AND JAN DE LEEUW

## 4.1 INTRODUCTION

A useful way to conceptualize ecological processes operating at different spatial scales is through what Wu (1999) calls hierarchical patch dynamics. A key notion is that a few parts of a large hierarchical structure can be studied in isolation insofar as these parts are distinguished from the rest by "near-decomposability." In effect, a segment of special interest within the hierarchical structure interacts weakly with the rest and then only asymmetrically. In this chapter, we focus on a particular kind of segment comprised of nested elements; higher levels are composed of the components of the level below. We consider multilevel statistical models that can be used to describe how variables characterizing higher levels affect processes operating at lower levels.

For simplicity, consider a subset of a hierarchy with two levels. The basic idea is to have a regression equation characterizing relationships at the lower (or micro) level and then have one or more of the regression coefficients at the micro level a function of predictors at the macro level. At the micro level, for instance, taxa richness may be a function of stream velocity (and other things). Then at the macro level, the regression coefficient linking stream velocity to taxa richness may be a function of proximity of the stream to land used for agriculture. Thus, one can address how the relationship between stream velocity and taxa richness varies (or not) in different locations, here with locale characterized by proximity to land use for agriculture. That is, one can learn when to generalize over sites and when not to generalize over sites. One can also learn how different scales are related.

These sorts of relationships can easily be formulated as interaction effects within a conventional regression analysis. However, the usual estimation procedures will not properly characterize the uncertainty in the output, so that the confidence intervals and hypothesis tests will not perform properly. A key problem is that the model's errors (or disturbances) are not likely to behave as if drawn independently from a single distribution. Special estimation procedures are required. Such



*J. Wu, K.B. Jones, H. Li, and O.L. Loucks (eds.),*
*Scaling and Uncertainty Analysis in Ecology: Methods and Applications*, 67–88.

procedures, often constructed within a multilevel framework, are well known and widely available in existing software (Raudenbush and Bryk, 2002). Our goal, therefore, is to summarize some recent extensions of multilevel models to more complicated and realistic situations common in ecological research. In the next section we provide an overview of the work. Technical details follow in subsequent sections (the more formal theoretical work here is primarily the work of de Leeuw).

## 4.2 EXTENSIONS OF MULTILEVEL MODELING

Our first extension of traditional multilevel modeling allows for spatial autocorrelation in the residuals of multilevel models. More proximate spatial units at the micro level can be expected to have model disturbances that are more alike than spatial units at the micro level that are more distant from one another. Thus, transects that are closer together will likely have disturbances that are more similar than transects that are farther apart. Failing to take this spatial autocorrelation into account will generally lead to biased estimates of the standard errors, and hence, inaccurate confidence intervals and hypothesis tests. Uncertainty will be characterized incorrectly.

Formally, a good solution to this problem for linear regression can be found in a classic paper by Ord (1975). For the usual sorts of regression models, one constructs a matrix capturing the distance between all micro units within each macro unit (e.g., transects within research sites) and builds that information into the estimation process. We initially adopted this approach, introduced it into a multilevel formulation, and applied it to two data sets. One data set was collected to study biodiversity in streams located in Ventura County, California, and the other was collected to study the impact of marine preserves on biodiversity and total fish biomass in coral reefs in the Philippines.

Our early results were disappointing. First, there was essentially no theory or empirical research in ecology or related disciplines to inform in sufficient detail the construction of the distance matrix. One difficulty was, for example, that it was not clear how to measure distance given ocean currents, which transport nutrients more readily between some locals than others. “Distance” was a function of spatial proximity and the direction and speed of prevailing current; locations formally closer together could easily have less in common than locations farther apart. Another difficulty was that there are a number of different distance functions that could have been used in the distance matrix (e.g., exponential decay with increasing distance) and, again, there is no guidance from the scientific literature. We believe that similar problems are common for a wide variety of environmental applications.

Second, except for very simple and somewhat unrealistic models, the numerical methods used in the estimation did not perform well. There were several technical reasons, but a key obstacle was that the regression coefficients and the distance matrix were “competing” for much the same information. This was because the predictors necessarily also contained spatial information. Micro units that were closer were likely to be more similar not just in their disturbances, but in the values of their predictors. Such predictors could include composition of the streambed and the amount of shading from trees along the banks, for instance. Because of the

competition for spatial information, the output from the statistical models tended to be very unstable. Small changes in the model or the data could introduce large changes in the output, which is a sure sign of trouble.

Finally, we planned to move beyond multilevel linear models to multilevel generalized linear models. In generalized linear models outcome variables can be counts or proportions. Thus, we would be able to include popular procedures, such as logistic regression for binary outcomes and Poisson regression for count data. Unfortunately, the Ord's approach led to effectively intractable mathematical problems when applied to generalized linear models.

These difficulties forced us to reconsider the entire enterprise and indeed, the usual philosophy by which spatial modeling is undertaken. To begin, we suspect that for spatial regression models, far too much is made about the exact form of the distance matrix. With scant scientific guidance about how the distance matrix should be formulated, any one of several competing formulations can be applicable. But, there is no way to know which is the best. In addition, the distance matrix by itself is rarely of much scientific interest. Its usual role is to allow for more accurate estimates of the regression coefficients that are the real focus of scientific concern. In statistical parlance, the distance matrix represents a set of "nuisance parameters."

At a deeper level, George Box's famous dictum applies: "all models are wrong, but some are useful." Given the current state of subject-matter knowledge, it is naive to aim for the "right" model. And in the absence of the right model, many of the usual statistical concerns become relatively unimportant. In particular, confidence intervals and tests no longer have much probative value. Rather, one should develop models that are informative and relatively simple and that capture in broad-brush strokes the essential features of the empirical world at hand (Berk 2003).

These and other considerations led us to consider methods by which the distance matrix could be well approximated and in a manner that eliminated much of the instability produced by taking the Ord approach. Two methods now seem to be especially effective. The first method extracts the eigenvectors of the distance matrix and uses the first few to adjust for spatial autocorrelation. This still requires, however, that a distance matrix be specified. The second method constructs simple functions of the spatial coordinates (e.g., longitude and latitude) and uses these to adjust for spatial autocorrelation. For example, one might include longitude, latitude, and their product. Analyses of real data and our own simulations indicate that both methods work well, although the second method is somewhat simpler to implement. Moreover, one can in both cases improve the approximation of the distance matrix as much as desired by using more of its eigenvectors or more complicated functions of the spatial coordinates. That is, one can make the approximations arbitrarily close to the specified distance matrix, although at some point the instabilities reappear. Finally, we have developed novel algorithms for estimating multilevel linear models with spatial autocorrelation that have been implemented in our software. The formal properties of these procedures have also been derived.

With our new approach, we can now easily turn to multilevel generalized linear models with spatial autocorrelation. All of the pieces are now in place. It is important to emphasize again, however, that we have in important ways reformulated the manner in which the modeling is approached; we are no longer

seeking the right model but rather, a useful model (For a rich elaboration on this point, see Berk 2003, p. 206-218).

## 4.3 THE FORMAL STRUCTURE OF MULTILEVEL MODELS

We have built on several existing traditions in statistics. Spatial regression models (Anselin 1988) are heteroscedastic linear models with correlated disturbances, in which the covariance between the disturbances depends on the spatial distance of the sites. Random coefficient models (Longford 1993) are heteroscedastic linear models with correlated disturbances, in which the covariance between the disturbances depends on the predictor similarity of the sites. Multilevel models (Kreft and de Leeuw 1998) are random coefficient models in which the predictor similarity is determined by the fact that sites are grouped into clusters. Disturbances between clusters are uncorrelated, but within clusters the covariance depends on the predictor similarity of the sites. Since distance and similarity are closely related constructs, one would expect a relationship between these three classes of models.

Spatial regression models and random coefficient models both have correlated disturbances, with the size of the correlation depending on the similarity of the sites. Similarity can be defined spatially or, more generally, in terms of similarity of the sites on a number of predictors which may not be spatial. Multilevel models simplify the overall correlation structure by assuming that sites in different clusters are uncorrelated, which means that the covariance matrix of the sites is block-diagonal, and presumably sparse.

It should come as no surprise, then, that much of our work relies on the earlier work of many others. But, we make four new contributions as follows: (1) we combine autoregressive models with multilevel models; (2) we consider spatial effects both as functions of non-spatial covariates with random coefficients and as autocorrelated disturbances; (3) we usefully approximate autocorrelated disturbance structures by using spatial regressors with random coefficients; and (4) we develop augmentation and majorization methods to estimate generalized multilevel autoregressive models. These are iterative computational methods dealing with the nonlinearities in such models (De Leeuw 1994).

The first three contributions are summarized below. Our work on the fourth contribution is available upon request. We also have a single, broad "take-home message." The development of statistical tools for environmental applications and the use of those tools should forego the traditional search for the "correct model" and focus instead on building one or more "useful models."

### *4.3.1 Basics*

We assume *multilevel data*. In the simplest case with two levels, the units of level one (which we call the *one-units*) are nested in units of level two (the *two-units*). Because of our concentration on spatial examples, we will often use the terminology of (research) "sites" and "transects" for the units in our levels. Transects are nested in sites. In the two-level case, we have $m$ two-units, and within two-unit $j$, we have

$n_j$ one-units. For each two-unit $j$, there is a vector $z_j$, of length $p$, of regressors describing the two-units. This implies that there will be $p$ regression coefficients, excluding the intercept, for two-units. There are also ($n_j$ x $q$) matrices $X_j$ of regressors describing one-units. This implies that there will be $q$ regression coefficients, excluding the intercept, for one-units. The total number of one-units in all $m$ two-units is $n$.

We usually allow for an intercept in regression models. So, we add a column of $n_j$ x 1 columns of 1's to $X_j$, with 1 as the lead element in $z_j$. Then, the standard two-level model assumes that within each two-unit $j$ we have a *random-coefficient regression model* of the form

$$\underline{y}_{ij} = \sum_{s=0}^{q} x_{ijs}\underline{\beta}_{js} + \underline{\varepsilon}_{ij}. \tag{4.1}$$

Here, $i$ is the index used for one-units ($I$ = 1, 2, …, $n_j$), which are nested in the two-units. We follow conventional practice and assume that the disturbances $\underline{\varepsilon}_{ij}$ are uncorrelated with the predictors; there are no "omitted variables" at the one-unit level, and the functional forms are appropriate. In practice, these assumptions must be carefully examined and justified. Often they will be found wanting. Note that random variables are always underlined and that we use element-wise notation initially, but matrix notation further on.

The *q+1 random regression coefficients* $\underline{\beta}_{js}$ in Equation 4.1 express the relationship between the *first-level predictors* and the *outcomes*. These random coefficients, of which there are p+1 for each two-unit $j$, are themselves outcomes of a second regression model, with fixed regression coefficients, shown in the equation

$$\underline{\beta}_{js} = \sum_{r=0}^{p} z_{jr}\gamma_{rs} + \underline{\delta}_{js}, \tag{4.2}$$

in which the random regression coefficients are outcomes predicted by *second-level predictors*. Again following convention, we assume that the disturbances, $\underline{\delta}_{js}$, are uncorrelated with the predictors; there are no "omitted variables" at the two-unit level, and the functional forms are appropriate. Of course, both assumptions are likely to be substantially wrong in practice, which again underscores the need to focus on useful models, not correct models. More will be said about this later.

In the spatial case, the first level predictors describe properties of the transects. They can be spatial, in the sense that they are functions of the coordinates of the transects, or non-spatial. The second level predictors describe properties of the sites, and again they can be spatial or non-spatial.

One can substitute Equation 4.2 into Equation 4.1 to write the model as a single equation:

$$\begin{aligned} \underline{y}_{ij} &= \Sigma_{s=0}^{q} x_{ijs} \{\Sigma_{r=0}^{p} z_{jr}\gamma_{rs} + \underline{\delta}_{js}\} + \underline{\varepsilon}_{ij} \\ &= \Sigma_{s=0}^{q}\Sigma_{r=0}^{p} x_{ijs} z_{jr}\gamma_{rs} + \Sigma_{s=1}^{q} x_{ijs}\underline{\delta}_{js} + \underline{\varepsilon}_{ij}. \end{aligned} \tag{4.3}$$

Thus, we see that the fixed part for two-unit $j$ has the form

$$E(\underline{y}_{ij}) = \sum_{r=0}^{p}\sum_{s=0}^{q} \gamma_{rs} z_{jr} x_{ijs} \tag{4.4}$$

with *(p+1)(q+1)* fixed predictors, each a product of a first-level and a second-level variable, often called "interaction variables." The random part has the form

$$\underline{y}_{ij} - \mathrm{E}(\underline{y}_{ij}) = \sum_{s=0}^{q} x_{ijs}\underline{\delta}_{js} + \underline{\varepsilon}_{ij}. \tag{4.5}$$

We now need some additional assumptions on the distribution of the disturbance terms. Some very general ones are: $\mathrm{E}(\underline{\varepsilon}_{ij}) = 0$; $\mathrm{E}(\underline{\delta}_{js}) = 0$; $C(\underline{\varepsilon}_{ij}, \underline{\varepsilon}_{k\ell}) = 0$ if $j \neq \ell$; $C(\delta_{js}, \delta_{\ell t}) = 0$ if $j \neq \ell$; and $C(\varepsilon_{ij}, \delta_{\ell s}) = 0$.

Thus, first-level disturbances for different two-units are uncorrelated, and so are second level disturbances. The dispersion matrices of the first-level disturbances are

$$\mathrm{E}(\underline{\varepsilon}_j \underline{\varepsilon}'_j) = \sigma_j^2 \Lambda_j, \tag{4.6}$$

and those of the second-level disturbances are

$$E(\underline{\delta}_j \underline{\delta}'_j) = \sigma_j^2 \Omega_j. \tag{4.7}$$

The dispersion matrix, $\sigma_j{}^2\Lambda_j$, allows the one-unit disturbances $\underline{\varepsilon}_{ij}$ for a given two-unit to have different variances and to be correlated with one another. The dispersion matrix, $\sigma_j^2\Omega_j$, allows the disturbances $\delta_{js}$ for a given two-unit to have different variances and to be correlated with one another. The former is where spatial dependence not captured by the regressors is likely to be seen. The latter will reflect dependence between the random coefficients that is not spatial, but a result of chance processes not captured by the two-unit model.

As a practical matter, it will be impossible to estimate the values of $\Lambda_j$ and $\Omega_j$. These matrices contain weights that determine the disturbance variances and covariances and as such, there are far too many parameters to estimate. Often to simplify we suppose that $\Omega_j$ are the same for all two-units, and usually $\sigma_j^2$ are supposed to be the same too. Still, in most cases (see the examples below), $\Omega_j$ and

$\Lambda_j$ are assumed to depend on a small number of parameters $\theta$, which may again be constant over two-units.

### *4.3.2 Example*

A simple spatial example may help clarify the model. It is not intended to be realistic, but to illustrate some key concepts. The one-units are observation stations, and the two-units are one of three counties. We suppose that rainfall at station *i* in county *j* depends on altitude (*alt*) and distance from the ocean (*dfo*).

$$\underline{rain}_{ij} = \underline{\beta}_{0j} 1_{ij} + \underline{\beta}_{1j} alt_{ij} + \underline{\beta}_{2j} dfo_{ij} + \underline{\varepsilon}_{ij}, \tag{4.8}$$

where $1_{ij}$ is the intercept, which is equal to one for all one-units. We do not assume that the regression coefficients are the same for all three counties. In fact, they vary according to a second regression model, for which we use indicator variables coding for the counties in the study. Thus, for s = 0, 1, 2,

$$\underline{\beta}_{js} = \gamma_{0s} 1_j + \gamma_{1s} LA_j + \gamma_{2s} SB_j + \underline{\delta}_{js}, \tag{4.9}$$

where again $1_{ij}$ is the intercept, now equal to one for all two-units. All observation stations in Los Angeles County (*LA*) have the same random coefficient distribution, and so do the observation stations in San Bernadino County (*SB*) and those in neither Los Angeles nor San Bernadino County.

If one substitutes the equations at the county level into the equations at the station level, for $i \neq k$ and assuming for notational simplicity that $\sigma_j^2$ and $\Omega_j$ are the same for all two-units,

$$C(\underline{rain}_{ij}, \underline{rain}_{kj}) = \sigma^2 [1 \quad alt_{ij} \quad dfo_{ij}] \begin{bmatrix} \omega_{00} & \omega_{01} & \omega_{02} \\ \omega_{10} & \omega_{11} & \omega_{12} \\ \omega_{20} & \omega_{21} & \omega_{22} \end{bmatrix} \begin{bmatrix} 1 \\ alt_{kj} \\ dfo_{kj} \end{bmatrix} \tag{4.10}$$

Thus, the covariance between the one-units in the same two-unit is determined by the similarity of predictor values of the one-units, where similarity is measured by their inner product in the matrix $\Omega$. This is a key insight, which shows why estimation of the parameters in spatial multilevel models can be difficult when one believes that certain sets of disturbances are correlated as well.

### *4.3.3 Matrix Notation*

Define the matrix $Z_j$ as the direct sum of $q$ copies of the row vector $z_j^t$. Thus, it is $q$ by $qp$, and it looks like

$$Z_j = \begin{bmatrix} z'_j & 0 & 0 & \dots & 0 \\ 0 & z'_j & 0 & \dots & 0 \\ 0 & 0 & z'_j & \dots & 0 \\ 0 & 0 & 0 & z'_j & 0 \\ 0 & 0 & 0 & 0 & z'_j \end{bmatrix} \tag{4.11}$$

Using this matrix and stacking the $\gamma_{rs}$ in a single vector $\gamma$, we can rewrite Equation 4.2 as

$$\underline{\beta}_j = Z_j \gamma + \underline{\delta}_j, \tag{4.12}$$

If we substitute Equation 4.12 into Equation 4.1, we find

$$\underline{y}_j = U_j \gamma + X_j \underline{\delta}_j + \underline{\varepsilon}_j, \tag{4.13}$$

$$E(\underline{y}_j) = U_j \gamma, \quad \text{with } U_j \triangleq X_j Z_j, \tag{4.14a}$$

$$V(\underline{y}) = \sigma_j^2 (X_j \Omega_j X'_j + \Lambda_j). \tag{4.14b}$$

It is convenient to write $\sum_j$ for $X_j \Omega_j X'_j + \Lambda_j$. Now $U_j$ is of the form

$$U_j = [x_{j1} z'_j \,|\, \dots \,|\, x_{jq} z'_j], \tag{4.15}$$

where $x_{jr}$ is column $r$ of $X_j$. Thus, in Equation 4.14a, the predictors in $U_j$ are products of a first-level predictor from $X$ and a second-level predictor from $Z$. In principle, all these *cross-level interactions* are part of the model, but we can eliminate some of them by setting the corresponding element of $\gamma$ equal to zero. Also observe that often the first column of both the $X_j$ and of $Z$ is an *intercept* column with all elements equal to +1. If we form all cross-level interactions, this implies that the columns of $X$ and $Z$ themselves also occur as predictors, because they are the intersections with the intercept at the other level.

### *4.3.4 Generalizations*

#### *4.3.4.1 More than two level*

In a more-than-two-level model, there are one-units, two-units, and so on, nested within each other. For instance, we can have transects nested within streams nested within watersheds. For this case we can adopt a more general notation. Suppose we have $n_r$ observations on level $r$, and $q_r$ predictors on that level. Thus, we have

$n_r \times (q_r + 1)$ matrices $X^{(r)}$ with predictors. We also use indicator matrices $G^{(r)}$, which are $n_r \times n_{r+1}$, and which indicate how the $r$-units map into the $(r+1)$-units.

The first two equations defining our multilevel model are

$$\underline{y}^{(1)}_{i_1} = \sum_{s_1=0}^{q_2} x^{(1)}_{i_1 s_1} \sum_{i_2=1}^{n_2} g^{(1)}_{i_1 i_2} \underline{y}^{(2)}_{i_2 s_1} + \underline{\varepsilon}^{(1)}_{i_1}, \tag{4.16a}$$

$$\underline{y}^{(2)}_{i_2 s_1} = \sum_{s_2=0}^{q_2} x^{(2)}_{i_2 s_2} \sum_{i_3=1}^{n_3} g^{(2)}_{i_2 i_3} \underline{y}^{(3)}_{i_3 s_1 s_2} + \underline{\varepsilon}^{(s)}_{i_2 s_1}. \tag{4.16b}$$

Thus, we have $n_1$ random variables in $\underline{y}^{(1)}$. These are the observed outcomes. We have $n_2 \times q_1$ unobserved random variables in $\underline{y}^{(2)}$, these are the random regression coefficients from our previous formulation. Then we have $n_3 \times q_1 \times q_2$ unobserved random coefficients in $\underline{y}^{(3)}$, and so on.

In the same way as before, we can combine equations to form single equations, which rapidly become unwieldy. For both mathematical reasons and ease of interpretation, it is wise to work with the fewest levels that can be justified. In practice, complex models also can have very unstable results and often will not converge at all. It is far better to have a model that is too simple than a model than is too complex. From Equation 4.16 we find, for example,

$$\underline{y}^{(1)}_{i_1} = \sum_{s_1=0}^{q_1} x^{(1)}_{i_1 s_1} \sum_{i_2=1}^{n_2} g^{(1)}_{i_1 i_2} [ \sum_{s_2=0}^{q_2} x^{(2)}_{i_2 s_1 s_2} \sum_{i_3=1}^{n_3} g^{(2)}_{i_2 i_3} y^{(3)}_{i_3 s_1 s_2} + \varepsilon^{(s)}_{i_2 s_1} ] + \underline{\varepsilon}^{(1)}_{i_1}. \tag{4.17}$$

### 4.*3.4.2 Multivariate outcomes*

If there is more than one outcome variable, we can use a simple trick to force the model into the multilevel framework. We use *variables* as the first level. Thus variables are nested in transects, transects in sites, and so on. For example, if there are three outcomes contained in three columns of the data set, one can reorganize the data so that within each one-unit there are three rows, one for each outcome. In each of these rows is the value for each of the three outcome variables respectively, with the values in the columns for predictor variables duplicated three times. Having multiple outcomes just adds a level to the hierarchy. In addition, missing data on the outcomes can be incorporated without difficulty, because some transects simply have fewer units (i.e., variables) than others.

We suspect that because there are often several ecologically interesting response variables for a given analysis, this approach to multiple outcomes can be widely useful. That is, even if there is no need for multilevel models because of a particular

hierarchical structure, multilevel level models can be used when there is a need to consider more than one outcome at a time.

#### *4.3.4.3 Non-independent two-units*

In our models, we usually assume that $\Omega_j$ are the same for all sites. With this assumption, it is possible to use a simple model for correlated sites, which has

$$C(\underline{y}_j, \underline{y}_\ell) = \sigma_{j\ell}(X_j \Omega X'_\ell + \Lambda_j^{1/2} \Lambda_\ell^{1/2} \tag{4.18a}$$

for all $j \neq \ell$, and

$$C(\underline{y}_j, \underline{y}_j) = \sigma_{jj}(X_j \Omega X'_j + \Lambda_j) \tag{4.18b}$$

for all $j$, where $\sigma_{j\ell}$ are the covariances between sites.

#### *4.3.4.4 Generalized MAR models*

In the same way as linear models are generalized to generalized linear models, one can try to construct generalized mixed linear models from mixed linear models. The trick is simply to condition on the random effects. In generalized linear models, first-level observations are independent given the random effects, and thus, the conditional distribution is a simple product of univariate Poisson, binomials, or gammas. But in generalized mixed linear models with autocorrelated or spatially correlated first-level disturbances, one no longer can use independence, and there is a need to assume that the disturbances within sites have multivariate Poisson, binomial, or gamma distributions. There is no agreement in statistics about how to define such multivariate distributions, and the definitions that are popular do not have many of the simplifying properties of the univariate versions.

We shall see below, however, that models with correlated first-level disturbances can be approximated by models with additional random effects and uncorrelated first-level disturbances. In these approximations, conditioning on the random effects makes the observations independent again, and the results developed for generalized mixed linear models apply again. This is perhaps the key technical point of this paper. To see why this works, we need to consider in greater detail the nature of the disturbances in the models.

## 4.4 MODELS FOR DISTURBANCE DISPERSIONS

The dispersion matrices $\Lambda_j$ of first-level disturbances can take many different forms. Generally, they are a function of a number of unknown parameters, collected in a vector $\rho$. Estimation simplifies considerably if the $\Lambda_j$ are known, and in particular in the homoscedastic case with uncorrelated disturbances in which

$\Lambda_j = I_j$, the identity matrix of order $n_j$. But in spatial situations the assumption that the disturbances are uncorrelated often is difficult to defend.

This is why a great deal of attention has been paid to modeling the dependence of spatial observations, taking as the main inspiration the literature on time series models. The key paper in spatial autoregressive (SA) modeling is Ord (1975). Also compare Griffith (2002b) and Anselin (2001). There are various forms of these SA models, but the most important ones are one-parameter models, in which the single parameter $\rho$ is interpreted as spatial autocorrelation. It indicates the strength of the spatial effects.

In multilevel models, restrictions are often placed on $\Omega_j$. For instance, it is common to assume that they are equal or that specific elements are zero. We shall discuss these restrictions later, and concentrate here on the first-level disturbances.

### *4.4.1 The Spatial Lag Model*

The spatial lag model is also known as the AR, or autoregressive response model. It specifies

$$\underline{y}_j = \rho_j W_j \underline{y}_j + X_j \underline{\beta}_j + \underline{\varepsilon}_j, \tag{4.19}$$

where $\underline{\varepsilon}_j$ is homoscedastic with variance $\sigma_j^2$. With $\underline{y}_j$ on both sides of the equal sign, this is an AR model with

$$E(\underline{y}_j \big| \beta_j) = (I_j - \rho_j W_j)^{-1} X_j \beta_j \tag{4,20a}$$

$$V(\underline{y}_j \big| \beta_j) = \sigma_j^2 [(I_j - \rho_j W_j)(I_j - \rho_j W_j')]^{-1}. \tag{4.20b}$$

In this formulation, the autoregression is defined directly in terms of the outcomes. The spatial dependence is built into the model in a structural manner. That is, the data analyst will typically have a subject-matter rationale for why and how values of the outcome variable are related. For example, if water quality in a lake is the outcome of interest, there may be diffusion of pollution from any one location to locations nearby. Depending on the value of $\rho_j$, the diffusion effects might be large or small, or perhaps even be negative. Note also that to isolate the role of the predictors, adjustments have to be made for the diffusion process, which links the outcome across locations. A failure to make such adjustments may mean that effects attributed to one or more of the predictors are really just a result of the movement of pollution from one place to another.

*4.4.2 The Spatial Error Model*

The spatial error model is also known as the SAR or simultaneous autoregressive model (Anselin 2001). It is expressed as

$$\underline{y}_j = X_j \underline{\beta}_j + \underline{\zeta}_j, \tag{4.21a}$$

and it assumes an autoregression structure for the error terms. Thus

$$\underline{\zeta}_j = \rho_j W_j \underline{\zeta}_j + \underline{\varepsilon}_j, \tag{4.21b}$$

where the $\underline{\zeta}_j$ are homoscedastic with variance $\sigma_j^2$. This leads to

$$E(\underline{y}_j \big| \beta_j) = X_j \beta_j, \tag{4.22a}$$

$$V(\underline{y}_j \big| \beta_j) = \sigma_j^2 [(I_j - \rho_j W_j)(I_j - \rho_j W_j')]^{-1}. \tag{4.22b}$$

This formulation implies that the spatial dependence is not potentially confounded with the predictors. It derives solely from dependence among the disturbances themselves. Disturbances that are more proximate in space, for instance, may tend to be more alike than disturbances that are farther apart. The reasons for the dependence are usually not of much interest. As such, the dependence is a mere nuisance and/or beyond current subject matter interest. The goal is to "mop up" the spatial dependence in the disturbances so that it does not affect the precision of estimates of the $\beta_j$ or estimates of their standard errors.

*4.4.3 The Conditional Autoregression Model*

Under the conditional autoregression model (CAR), also discussed in Anselin (2001), we let

$$\underline{y}_j = X_j \underline{\beta}_j + (I_j - \rho_j W_j)^{-1/2} \underline{\varepsilon}_j, \tag{4.23}$$

where $W_j$ is now a symmetric weight matrix, and $\underline{\varepsilon}_j$ is homoscedastic with variance $\sigma_j^2$. This implies

$$E(\underline{y}_j \big| \beta_j) = X_j \beta_j, \tag{4.24a}$$

$$V(\underline{y}_j \big| \beta_j) = \sigma_j^2 [(I_j - \rho_j W_j)]^{-1}. \tag{4.24b}$$

If dependence in the disturbances can be treated as a mere nuisance, the model that one uses for the disturbances is of little importance as long as the dependence is taken into account when the regression coefficients are estimated. In this context, the conditional autoregressive model can be seen as an alternative to the spatial error model, and it has some of the same look and feel. Larger values of $\rho_j$ imply more dependence among the disturbances. And just as for the spatial error model, the dependence may be a function of distance; closer disturbances may tend to be more alike. The main advantage of CAR model is that it can be as effective in mopping up dependence in the disturbances as the spatial errors model, but will be far easier to compute.

### *4.4.4 Weight Matrices*

How to choose $W_j$ has been discussed many times in the geostatistics literature. A good review is Bavaud (1998; see also Cressie 1991). Although it is possible to give some general indications, choosing a precise and appropriate $W_j$ is difficult, probably even more difficult than choosing a correct set of predictors. The usual problem is that there is too little a priori knowledge to inform the choice and at best some general clues in the data.

#### *4.4.4.1 Choice of weights*

For $W_j$ in spatial situations, we assume that its elements are similarities of transects in site $j$. The more similar (the closer) the transects, the larger the corresponding element in $W_j$. If we do not have a good reason to choose a specific $W_j$, we can make it some (decreasing) function of the transect distances, but again choosing the function is often disturbingly arbitrary. In many cases, moreover, we even want to replace simple Euclidean distance by other distances (measured along a network or stream, for instance), which take the actual spatial setting into account. Throughout, we suppose the elements of $W_j$ are non-negative.

#### *4.4.4.2 Large matrices*

In spatial analysis we often encounter situations in which the order of $W_j$ is very large, maybe $10^5$ or $10^6$. Obviously in such cases, it will generally not be possible to store floating-point matrices of this size, let alone compute their determinants, inverse, or eigen-decomposition.

There are several ways around this problem. The first is to use patterned weight matrices of zeroes and ones (coding adjacency or nearest neighbor, for instance), with a determinant or an inverse available in analytical form (Pace and Zou 2000). The second is to use sparse matrix techniques for weight matrices with a very large proportion of zeroes (Pace and Barry 1997a, 1997b, 1997c) (again, adjacency matrices come to mind). We have also seen that multilevel analysis suggests partitioning transects or sites into clusters, and making the between cluster covariance equal to zero. This also introduces a great deal of sparseness. And finally, fast

numerical approximations to the loss function are also a possibility. Specifically, techniques for approximating the determinant in the normal log-likelihood for all AR, SAR, and CAR models are in Smirnov and Anselin (2001) and Griffith (2002a).

In the models discussed in this paper, we have the additional complication that the dispersion matrix is made up of two components: a part based on similarity of the regressors and a part based on spatial information, coded in the weight matrices. This makes patterned weight matrix and sparse matrix techniques more difficult to use, and we have to resort to other types of approximations.

*4.4.4.3 Normalizing the weights*

It is computationally convenient if the weight matrices in the SAR and AR models are symmetric. Then, we get a more simple formulation,

$$(I_j - \rho_j W_j)(I_j - \rho_j W_j') = (I_j - \rho_j W_j)^2,$$

which is easier to work with. Unfortunately, in many applications an asymmetric set of weights may make more sense (think of the influence of stream flow or hillside slope on ecological distance, for instance).

Consider why having symmetric matrices is convenient. If the $W_j$ are known symmetric matrices, one can compute the spectral decomposition, $W_j = K_j \Phi_j K_j'$, and we find

$$\Lambda_j(\rho_j) = \sum_s \frac{1}{(1-\rho_j \phi_{js})^2} k_{js} k_{js}' \text{ for SAR} \tag{4.25a}$$

$$\text{and } \Lambda_j(\rho_j) = \sum_s \frac{1}{1-\rho_j \phi_{js}} k_{js} k_{js}' \text{ for CAR.} \tag{4.25b}$$

Thus, the eigenvectors of $\Lambda_j(\rho_j)$ are the same as those of $W_j$, and the eigenvalues are simple functions of the eigenvalues of $W_j$. If $\rho_j$ changes, only the eigenvalues change; the eigenvectors remain the same.

For interpretation purposes, one can normalize the weights in such a way that the rows of $W_j$ sum to unity. This makes the weight matrix stochastic, and by Frobenius theorem implies that the largest eigenvalue of $W_j$ is equal to +1. This means that the smallest eigenvalue of $I_j - \rho_j W_j$ is $1 - \rho_j$, and thus $I_j - \rho_j W_j$ is positive definite as long as $\rho_j < 1$, which helps in the interpretation of $\rho$ as a type of autocorrelation coefficient.

In some cases, it is desirable for $W_j$ to be both symmetric and normalized (i.e., doubly stochastic). This is discussed for CAR models in Page and LeSage (2002). We have developed an algorithm and computer codes to normalize non-negative

symmetric matrices in such a way that they become doubly stochastic (the codes available upon request).

### *4.4.5 Special Case: Time Series Models*

If the outcomes are one-dimensional (e.g., if transects are arranged in lines), then it makes sense to use a time series model for the first-level disturbances (Hedeker 1989, Hedeker and Gibbons 1996). We discuss these models here briefly because they show where the SA models come from, and because they are more likely to be familiar.

A first obvious choice for a time-series model is the *random walk*, which has

$$\underline{\varepsilon}_j = W_j \underline{\varepsilon}_j + \underline{\zeta}_j, \tag{4.26}$$

where $W_j$ has all elements equal to zero, except for those immediately below the main diagonal, which are one. It follows that

$$\underline{\varepsilon}_j = T_j \underline{\zeta}_j, \tag{4.27}$$

where $T_j$ has all elements on and below the main diagonal equal to one and all elements above the main diagonal equal to zero. Thus,

$$\Lambda_j = T_j T_j', \tag{4.28}$$

which means that element ($s$, $t$) is equal to *min(s, t)*.

In an AR(p) process,

$$\underline{\varepsilon}_j = W_j \underline{\varepsilon}_j + \underline{\zeta}_j, \tag{4.29}$$

where $W_j$ has a band of width $p$ below the diagonal and zeroes elsewhere. There are $p$ parameters, the autoregression coefficients, in $W_j$. The AR(1) model is much like the random walk, except that the element below the diagonal is the single parameter $\rho_j$.

An MA($q$) process also uses a banded matrix with parameter values, but now

$$\underline{\varepsilon}_j = W_j \underline{\zeta}_j, \tag{4.30}$$

where $W_j$ has diagonal one, and a band of width $q$ in each row below the diagonal. Thus, an MA(1) has diagonal one, and $\rho_j$ below the diagonal.

It is easy to extend this to ARMA($p, q$) and even more complicated processes, but this is comparatively straightforward and it may be overkill in many practical situations. For our purposes, the most interesting models are AR(1) and MA(1),

which can be defined in term of the backshift matrix $B_j$, which has elements equal to one below the diagonal only. Then for AR(1), we have

$$\Lambda_j(\rho_j) = (I_j + \rho_j B_j)^{-1}(I_j + \rho_j B_j')^{-1}, \tag{4.31a}$$

and for MA(1) we have

$$\Lambda_j(\rho_j) = (I_j + \rho_j B_j)(I_j + \rho_j B_j'). \tag{4.31b}$$

The random walk is AR(1) with $\rho_j = 1$.

## 4.5 MODEL APPROXIMATION

Now we turn to approximations of various AR models that can lead to practical computational results. Because, as noted earlier, the concept of "the true model" is at least obscure and because even if we know how to think about "the true model," we usually do not have very precise information about which $W_j$ produces it, it makes sense to employ an approximation of the dispersion matrix that is computationally convenient. We will first simplify the model by an approximation that works well for small $\rho_j$, and then we approximate the model by another formulation with homoscedastic first-level disturbances (i.e., a model with $\Lambda_j = I_j$).

### *4.5.1 Simplified AR*

Consider again the SAR model described in the section of the spatial error model above. Recall that the variance-covariance matrix of the disturbances was $\sigma_j^2[(I_j - \rho_j W_j)(I_j - \rho_j W_j)]^{-1}$, where all of the terms to the right of $\sigma_j^2$ represent $\Lambda_j(\theta)$. In the *Simplified* AR Model (SIMAR), assume

$$\Lambda_j(\theta) = I_j + \rho_j W_j, \tag{4.32}$$

where the off-diagonal elements of the symmetric matrix $W_j$ are again some decreasing function of the Euclidean distances between the transects or, more generally, of the spatial dissimilarities.

In the CAR model, if $\rho_j$ is small,

$$(I_j - \rho_j W_j')^{-1} = I_j + \rho_j W_j + o(\rho_j), \tag{4.33}$$

and in the SAR and CAR models,

$$\begin{aligned}\Lambda_j(\rho_j) &= (I_j - \rho_j W_j)^{-1}(I_j - \rho_j W_j')^{-1} \\ &= I_j + \rho_j(W_j + W_j') + o(\rho_j),\end{aligned} \tag{4.34}$$

which are both of the SIMAR form.

For both AR(1) and MA(1), and small $\rho_j$,

$$\Lambda_j = I_j + \rho_j(B_j + B_j') + o(\rho_j), \tag{4.35}$$

which is again of the required SIMAR form.

*4.5.2 Spatial Effects As Random Coefficients*

By using random coefficients in appropriate ways, one can emulate the covariance structure of the SIMAR without assuming correlated disturbances for the first-level units. Thus, one can maintain $\Lambda_j = I_j$. The trick is really quite simple. In our spatial multilevel models

$$\underline{y}_j = U_j\gamma + X_j\underline{\delta}_j + \underline{\varepsilon}_j, \tag{4.36}$$

$$V(\underline{\varepsilon}_j) = \sigma_j^2(I_j + \rho_j W_j). \tag{4.37}$$

Now suppose $W_j = K_j\Phi_j K_j'$ is the spectral decomposition of $W_j$. Then,

$$\underline{y}_j = U_j\gamma + X_j\underline{\delta}_j + K_j\underline{\eta}_j + \underline{\zeta}_j, \tag{4.38}$$

where $\underline{\delta}_j$ and $\underline{\eta}_j$ are uncorrelated, and where

$$V(\underline{\eta}_j) = \sigma_j^2\rho_j\Phi_j, \tag{4.39a}$$

$$V(\underline{\zeta}_j) = \sigma_j^2 I_j. \tag{4.39b}$$

But, Equations 4.38 and 4.39 can be interpreted as a simple multilevel model in which the covariance matrix of the random effects is of the form

$$\begin{bmatrix}\Omega & 0 \\ 0 & \rho_j\Phi_j\end{bmatrix}. \tag{4.40}$$

First-level disturbances are homoscedastic, and the regression coefficients corresponding with the eigenvector-predictors $K_j$ only have a random part and a

vanishing fixed part. Moreover, the random parts are uncorrelated, with a diagonal dispersion matrix proportional to the eigenvalues of $W_j$. In short, one can write the SIMAR model as a multilevel model with restrictions on the covariance matrix of the random effects.

### *4.5.3 Positive Definite Variances*

One problem with this formulation is that it is not guaranteed that the eigenvalues $\Phi_j$ of $W_j$ are non-negative. If there are negative eigenvalues, then Equation 4.39a is difficult to interpret. One can use the fact, however, that $I_j + \rho_j W_j$ must be positive definite. Suppose $\rho_j > 0$, and write $y_j$ for the smallest eigenvalue of $W_j$. Then,

$$I_j + \rho_j W_j = (1 + \rho_j \psi_j) I_j + \rho_j K_j (\Phi_j - \psi_j I_j) K_j', \tag{4.41}$$

and we can rewrite Equation 4.39 as

$$V(\underline{\eta}_j) = \sigma_j^2 \rho_j (\Phi_j - \psi_j I_j), \tag{4.42a}$$

$$V(\underline{\zeta}_j) = \sigma_j^2 (1 + \rho_j \psi_j) I_j. \tag{4.42b}$$

These are somewhat more complicated restrictions, but they always give positive semidefinite dispersion matrices.

### *4.5.4 Using Fewer Eigenvalues*

A second problem with our approximation is that we replace a very large spatial disturbance covariance matrix with a very large number of random effects. The number of random effects added is equal to the order of the spatial covariance matrix. We attack this problem by using only a small number of eigenvectors of $W_j$, those corresponding with the largest eigenvalues (in modulus). Thus, we use a principal component type approximation to the random effects. With spatial information in $W_j$ using some function of the distances, two or three principal components are likely to give a rather good approximation.

### *4.5.5 General Approach*

Instead of approximating the SA models by SIMAR, and then approximating SIMAR by using eigenvectors, one can employ a more straightforward approach that can reduce the computational burdens. Consider the following multilevel model for site *j*

$$\underline{y}_j = X_j \underline{\beta}_j + Z_j \eta_j + \underline{\varepsilon}_j, \tag{4.43}$$

where $X_j$ contains regression coordinates, and $Z_j$ contains functions of the spatial coordinates. For our second level model, we use

$$\underline{\beta}_j = A_j\gamma + \underline{\delta}_j, \tag{4.44a}$$

$$\underline{\eta}_j = B_j\kappa + \underline{\xi}_j. \tag{4.44b}$$

This implies

$$\underline{y}_j = X_jA_j\gamma + Z_jB_j\kappa + \underline{\nu}_j, \tag{4.45a}$$

where

$$\underline{\nu}_j = X_j\underline{\delta}_j + Z_j\underline{\xi}_j + \underline{\varepsilon}_j, \tag{4.45b}$$

and thus, with suitable uncorrelatedness assumptions,

$$E(\underline{y}_j) = X_jA_j\gamma + Z_jB_j\kappa, \tag{4.46a}$$

$$V(\underline{y}_j) = \sigma_j^2(X_j\Omega_jX_j' + Z_j\Theta_jZ_j' + I_j). \tag{4.46b}$$

This becomes an approximate multilevel Ord model if we let $B_j = 0$ (i.e., the spatial regression coefficients do not have a fixed part), and we let $\Theta = \rho_j^2 I_j$ (i.e., the spatial regression coefficients are uncorrelated). Then,

$$E(\underline{y}_j) = X_jA_j\gamma, \tag{4.47a}$$

$$V(\underline{y}_j) = \sigma_j^2[X_j\Omega_jX_j' + (I_j + \rho_j^2Z_jZ_j')]. \tag{4.47b}$$

Moreover, to get closer to SA, one can choose $Z_j$ in clever ways, using the results that we discussed earlier in this section. If the $W_j$ matrix in the Ord model is a function of the spatial distances, then it obviously is a function of the coordinates, and thus all its eigenvectors are functions of the coordinates. If we choose $Z_j$ as a low-rank (principal component) approximation of $W_j$, using the eigenvectors, then we can get very close to the Ord model.

With the practical approximation of the Ord Model, applications to the generalized linear model follow easily and directly. Work on these extensions is nearly completed, and software development has begun.

## 4.6 EMPIRICAL EXAMPLE

We have data from coral reefs along Olango Island in the Phillipines (unpublished data provided by Craig Schuman and Domingo Ochavillo). There are 33 sites with four transects in each. There are 14 sites in areas that are protected; fishing is prohibited. There are 19 sites that are in unprotected areas; fishing is allowed (and is common). And the fishing can include very destructive practices such as poisoning fish. The data that we analyze is an aggregate over four equally spaced observations along each transect. Thus, transects are our lowest level, and the second level is sites in the multilevel spatial model.

To keep the example simple, we use the same formulation illustrated in the example section above. The main difference is that the specific response is the number of different fish species. Therefore, the results reported are for a Poisson response variable within our multilevel framework. At the lowest level, the number of fish species is a function of how sandy the bottom is. The larger the percent of the bottom that is sandy, the fewer species one would expect. This relation depends on the intercept ( $\beta_0$) and the slope of regression ( $\beta_1$) at the level of the site. Then, the intercept is a function of whether the reef is protected via $\gamma_0$. Note that a reef is coded 1 if the reef is protected and 0 otherwise. The slope is also a function of whether the reef is protected via $\gamma_1$. Thus, both of the parameters at the level of the site are treated as random coefficients with a structural component determined by whether the reef is protected. In Table 4.1, we used an augmentation algorithm, but the results are much the same for any of the simplifications that we have discussed.

***Table 4.1.*** *Model for species counts estimated by augmentation algorithm (N = 132).*

| *Predictor* | *Coefficient* | *Standard Error* |
|---|---|---|
| Protected ( $\gamma_0$) | 5.58 | 3.75 |
| % Sandy Bottom ( $\beta_1$) | -0.18 | 0.04 |
| Protect *X* % Sandy ( $\gamma_1$) | 0.05 | 0.08 |
| Constant ( $\beta_0$) | 27.6 | 2.10 |
| $\theta$ (AR parameter) | 0.44 | – |

Focusing first on the regression coefficients from Table 4.1, one can see that if a reef is unprotected there are on average nearly 28 distinct fish species at a site. At these unprotected sites, for each additional percent of the bottom that is sandy, the number of species drops by 0.18; for every additional 10%, the number of species drops by 1.8. In the protected sites, the number of fish species is greater by 5.58. Finally, in the protected sites, the negative impact of a sandy bottom on the number of species is a bit less pronounced. The regression coefficient of -0.18 is now -0.13. For every 10% increase in sandy bottom, the number of species is reduced by 1.3.

The autoregressive parameter is 0.44, which is of moderate size. There is some meaningful spatial autocorrelation in the residuals. When this is taken into account, we see in Table 4.1 that the percent of the bottom that is sandy is easily twice the

standard error. The impact of protecting a reef is about 1.5 times its standard error, statistically significant at the 0.10 level for a one-tailed test. The coefficient for the interaction effect is less than its standard error. One should treat any formal tests with great caution, in part because the data were not collected by random sampling, and there is no compelling model-based sampling alternative. Also, it is virtually certain that important explanatory variables have been overlooked. But if one chooses to take formal tests seriously, the interaction effect can be discarded.

## 4.7 SOFTWARE

The results shown in Table 4.1 were produced by developmental software that we wrote in R. But, one can obtain consistent estimates of all the regression parameters using any software for Poisson regression, usually as a special case of the generalized linear model. One just has to substitute the higher level equations into the lower level equations, simplify, and proceed as usual. And if confidence intervals and tests are not formally justified (which is often the case), one does not have to proceed any farther. Moreover, all of the conventional regression diagnostics for the generalized linear model apply (Cook and Weisberg 1999).

Getting the uncertainty right is more difficult and, at this point, requires special software. We hope to have ours available soon. In the meantime, there are two good options that can be employed with existing software. First, if one can justify the spatial lag model discussed above, then once the substitution of the two-unit model into the one-unit model is completed, the GLM version of Equation 4.19 (e.g., logistic regression or Poisson regression) can be estimated in all of the major statistical packages with their routine GLM procedures. Consistent parameter estimation follows. Second, and far more generally, there exists, at least in SAS PROC MIXED (Littell et al. 1996), MLwiN (Goldstein 1995), GLLAMM (Rabe-Hesketh et al. 2002), and HLM (Raudenbush and Bryk 2002), the ability to do mixed effects generalized linear models. For a comparison of these packages we refer to De Leeuw and Kreft (2001). In each of them one can include functions of the spatial coordinates for the one-units but with their regression coefficients constrained to be equal to zero. We have had good success including just the horizontal coordinate, the vertical coordinate (e.g., longitude and latitude), and their product as one-unit predictors. Most of the spatial autocorrelation in the model will likely be "soaked-up." The standard errors should then be sufficiently accurate for most purposes.

## 4.8 CONCLUSIONS

In this paper, we discussed tools for the construction of multilevel linear models with ecological data. Extensions to multilevel generalized linear models followed directly. With these tools, one can examine how variables at one level are related to processes at another level; one can study the interactions between phenomena at different spatial/temporal scales. If one can also make the case that the structure of a model is very nearly right, and one has either random sampling or credible-model

based sampling (Berk, 2003), conventional ways of representing uncertainty apply. Our suggestions for obtaining useful estimates of the standard errors are then appropriate. However, we favor a more realistic approach in which description is the primary goal.

## ACKNOWLEDGEMENTS

This work was supported by a grant from the EPA STAR program.

## REFERENCES

Anselin, L. 1988. Model validation in spatial econometrics: a review and evaluation of alternative approaches. International Regional Science Review 11:279-316.

Anselin, L. 2001. Spatial Regression. http://geog55.gis.uiuc.edu/~luc/talks/spreg.pdf.

Bavaud, F. 1998. Models for spatial weights: a systematic look. Geographical Analysis 30:153-171.

Berk, R. 2003. Regression Analysis: A Constructive Critique. Sage Publications, Newbury Park, CA.

Cook, R. D., and S. Weisberg. 1999. Applied Regression Including Computing and Graphics. Wiley.

Cressie, N. 1991. Statistics for Spatial Data. Wiley.

de Leeuw, J. 1994. Block relaxation methods in statistics. Pages 275-289 in H. H. Bock, W. Lenski, and M. M. Richter, editors. Information Systems and Data Analysis. Springer, New York.

de Leeuw, J., and I. G. G. Kreft. 2001. Software for multilevel analysis. Pages 187-204 in A. H. Leyland and H. Goldstein, editors. Multilevel Modelling of Health Statistics. Wiley, New York.

Goldstein, H. 1995. Multilevel Statistical Medels. Arnold.

Griffith, D. A. 2002a. Quick but not so Dirty. ML Estimation of Spatial Autoregressive Models. Technical Report. Department of Geography, Syracuse University, Syracuse, NY.

Griffith, D. A. 2002b. Spatial Autoregression. Syracuse University, Syracuse, NY.

Hedeker, D. R. 1989. Random Regression Models with Autocorrelated Errors: Investigating Drug Plasma Levels and Clinical Response. Ph.D. Thesis. University of Chicago at Illinois, Chicago, IL.

Hedeker, D. R., and R. D. Gibbons. 1996. MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. Computer Methods and Programs in Biomedicine 49:229-252.

Kreft, I., and J. de Leeuw. 1998. Introducing Multilevel Modeling. Sage Publications, Newbury Park, CA.

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. SAS System for Mixed Models. SAS Institute, Cary, NC.

Longford, N. 1993. Random Coefficient Models. Oxford University Press, Oxford.

Ord, J. K. 1975. Estimation methods for models of spatial interactions. Journal of the American Statistical Association 70:120-127.

Pace, R. K., and R. Barry. 1997a. Fast CARs. Journal of Statistical Compution and Simulation 59:123-147.

Pace, R. K., and R. Barry. 1997b. Quick computation of spatial autoregressive estimators. Geographical Analysis 29:232-246.

Pace, R. K., and R. Barry. 1997c. Sparse spatial autoregressions. Statistics and Probability Letters 33:191-197.

Pace, R. K., and D. Zou. 2000. Closed-form maximum likelihood estimates of nearest neighbor spatial dependence. Geographic Analysis 32:154-172.

Page, R. K., and J. P. LeSage. 2002. Conditional Autoregressions with Doubly Stochastic Weight Matrices (manuscript).

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal 2:1-21.

Raudenbush, S. W., and A. S. Bryk. 2002. Hierarchical Linear Models, second edition, second edition. Sage Publications, Newbury Park, CA.

Smirnov, O., and A. Anselin. 2001. Fast maximum likelihood estimation of very large spatial auto-regressive models. Computational Statistics and Data Analysis 35:301-319.

Wu, J. 1999. Hierarchy and scaling: extrapolating information along a scaling ladder. Canadian Journal of Remote Sensing 25:367-380.