Ecological Applications, 16(1), 2006, pp. 20–32

STATISTICS FOR CORRELATED DATA: PHYLOGENIES, SPACE, AND TIME

ANTHONY R. IVES[1,3] AND JUN ZHU[2]

[1]*Department of Zoology, University of Wisconsin–Madison, Madison, Wisconsin 53706 USA*
[2]*Department of Statistics, University of Wisconsin–Madison, Madison, Wisconsin 53706 USA*

Abstract. Here we give an introduction to the growing number of statistical techniques for analyzing data that are not independent realizations of the same sampling process—in other words, correlated data. We focus on regression problems, in which the value of a given variable depends linearly on the value of another variable. To illustrate different types of processes leading to correlated data, we analyze four simulated examples representing diverse problems arising in ecological studies. The first example is a comparison among species to determine the relationship between home-range area and body size; because species are phylogenetically related, they do not represent independent samples. The second example addresses spatial variation in net primary production and how this might be affected by soil nitrogen; because nearby locations are likely to have similar net primary productivity for reasons other than soil nitrogen, spatial correlation is likely. In the third example, we consider a time-series model to ask whether the decrease in density of a butterfly species is the result of decreases in its host-plant density; because the population density of a species in one generation is likely to affect the density in the following generation, time-series data are often correlated. The fourth example combines both spatial and temporal correlation in an experiment in which prey densities are manipulated to determine the response of predators to their food supply.

For each of these examples, we use a different statistical approach for analyzing models of correlated data. Our goal is to give an overview of conceptual issues surrounding correlated data, rather than a detailed tutorial in how to apply different statistical techniques. By dispelling some of the mystery behind correlated data, we hope to encourage ecologists to learn about statistics that could be useful in their own work. Although at first encounter these techniques might seem complicated, they have the power to simplify ecological research by making more types of data and experimental designs open to statistical evaluation.

Key words: comparative methods; correlated data; geostatistical model; mixed model; phylogenetic correlation; statistical methods; time-series analysis.

INTRODUCTION

The first line in the description of many common statistical tests is that the data represent independent samples from the same statistical distribution; the sampled observations must be identically and independently distributed (iid). This often poses a problem for ecologists, particularly those with observational (nonexperimental) data taken from similar species, in similar locations, or at similar points in time. There is often no guarantee that ecologists' data represent independent samples, and, in fact, there is often strong reason to believe they do not. Even in experimental studies, practical constraints may make randomized designs impossible. This may lead to pseudoreplication, in which treatment replicates are not independent (Hurlbert 1984). The need for independence, and the fear that nonindependence generates, potentially limit the types of studies that ecologists perform.

The limitations imposed by the need for independence are being overcome, and there are several statistical approaches that are now commonly used by ecologists that treat nonindependent, or correlated, data. Perhaps the most familiar are repeated-measures analyses (Snedecor and Cochran 1989: chapter 16.6, Neter et al. 1996: chapter 29), especially repeated-measures ANOVA, in which multiple samples are taken from the same individual, plot, population, etc. Because measurements taken repeatedly from the same unit are likely to be more similar to each other than measurements from different units, repeated-measures analyses explicitly account for correlation in the data taken from a single unit. Analyses of time-series data also account for correlated data, because measurements taken close in time may be more similar than measurements separated by long periods of time (Box et al. 1994). Like

Manuscript received 22 April 2004; revised 19 August 2004; accepted 24 August 2004; final version received 21 October 2004. Corresponding Editor: D. S. Schimel. For reprints of this Invited Feature, see footnote 1, p. 3.

[3] E-mail: arives@wisc.edu

TABLE 1. Sources of correlation and approaches to estimation and statistical inference discussed in the article.

Source of correlation	Estimation	Inference
Phylogeny	generalized least squares (GLS)	exact computation
Space	maximum likelihood (ML)	asymptotic approximation
Time	estimated generalized least squares (EGLS)	parametric bootstrapping
Mixed space and time	restricted maximum likelihood (REML)	asymptotic approximation

time, space also generates correlation, as samples taken from nearby locations may be more similar to each other than samples taken far apart due to some unmeasured factor that varies spatially (Cressie 1991). Finally, linear mixed models that include both planned ''fixed'' effects and unplanned ''random'' effects (Snedecor and Cochran 1989: chapter 13.9, Littell et al. 1996, Neter et al. 1996, chapter 24) are seeing greater use in ecology. The flexibility of mixed models allows them to incorporate numerous types of correlation in data sets.

This article gives a general presentation of the consequences of correlated data and how it can be treated statistically. It is not intended as a tutorial that gives step-by-step instructions on how to perform statistical tests, nor is it a flow chart to guide readers to a particular method needed for some data set in hand. Instead, we intend this article to be a companion to books and manuals that address specific techniques. This article gives a general discussion of how correlation is built into statistical methods, and how correlation arising in many different ways can nonetheless be analyzed with similar methods.

The problems we address all have the general structure of regression

$$y_i = b_0 + b_1 x_i + \varepsilon_i \quad (1)$$

where x is the independent variable, y is the dependent variable, and the i denotes the ith data sample. (Owing to an unfortunate lexicon, the designation of variables as independent or dependent is unrelated to the independence or nonindependence of the samples). The error terms ε_i designate unexplained variability, and b_0 and b_1 are regression coefficients. For standard regression, the error terms ε_i are normally and independently distributed. Here we consider the case in which error terms are normally but not independently distributed.

The statement and analyses of a general regression problem can be done most easily in matrix notation. We assume that the reader has some knowledge of matrix algebra; if not, Neter et al. (1989) give a good introduction. In matrix notation, the regression problem can be stated as

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon} \quad (2a)$$

$$\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{V} \quad (2b)$$

where $\mathbf{X}$ is the matrix whose first column contains ones and second column contains values x_i, $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are column vectors of values y_i and ε_i, $\mathbf{b}$ is the vector $(b_0, b_1)'$, and values of $\boldsymbol{\varepsilon}$ follow a multivariate normal distribution with means 0 and covariance matrix $\sigma^2\mathbf{V}$. The covariance matrix is typically written in the form $\sigma^2\mathbf{V}$, where σ^2 determines the overall magnitude of the variance and $\mathbf{V}$ gives the correlation structure of the data. The element in the ith row and jth column of the covariance matrix, $\sigma^2 v_{ij}$, contains the covariance between the error terms ε_i and ε_j. In a standard regression problem, $\mathbf{V}$ is the identity matrix $\mathbf{I}$, but this is just a special case.

We consider four scenarios that generate correlated data. While on the surface these scenarios might appear quite different, they are in fact quite similar. Although we want to emphasize their similar structure, we will use each scenario to illustrate different statistical approaches for obtaining values of model parameters (estimation) and statistical confidence and tests of those values (inference). Table 1 lists the four scenarios and the approaches to estimation and inference we consider. There are additional ways in which correlated data may arise and additional approaches to estimation and inference. Nonetheless, this list gives a starting point for our introduction to statistics for correlated data. We discuss each scenario using a simple, hypothetical example of a type that might arise in ecological studies. To generate data for these examples, we simulate the same model that we then use to fit the data statistically. Therefore, we do not address the issues of model identification/selection and model validation (Neter et al. 1989, Burnham and Anderson 1998). Computer codes for our illustrative examples are given in the Supplement, and Appendix A gives a list of useful references.

PHYLOGENETIC DEPENDENCE IN COMPARISONS AMONG SPECIES

Example problem 1

Suppose you collected data on the body size (mass) and home-range area for 15 ungulate species, and you wanted to determine whether larger animals have larger home ranges (Garland et al. 1992). This is a regression problem in the form of Eq. 2 in which y = home range area and x = body size (where both variables may have been transformed). The challenge for the analysis, however, is that species do not represent independent samples from an underlying distribution of home-range area given body size (Felsenstein 1985). This is because more closely related species tend to be more

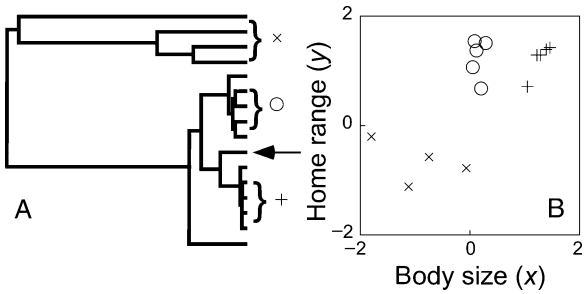

FIG. 1. Regression analysis for phylogenetically correlated data: (A) hypothetical phylogenetic tree for 16 species; (B) simulated values of home-range area, *y*, and body size, *x*, for 15 of 16 species. Points shown as ×, ○, and + correspond to the clades marked by the same symbols in panel A. The value of home-range area is unknown for the species shown by the arrow in panel A. Values of body size were simulated to conform to the correlation predicted from the phylogenetic tree, and values of home-range area were simulated from the regression model given by the regression Eq. 2 with parameter values given in Table 2.

similar (Blomberg et al. 2003). In this simulated example, the signature of phylogeny is seen in the clustering of data, in which three lineages, or clades (marked with different symbols in Fig. 1), have similar values of both home-range area and body size. Thus, phylogeny produces correlation among data from related species, and this correlation must be factored into the statistical analysis (Felsenstein 1985, Garland et al. 1992).

To account for the phylogenetic correlations among species, it is necessary to translate the phylogenetic tree into the covariance matrix of the error terms ε in the regression model (Eq. 2). To do this, first consider the evolution of a single trait from the base to the tips of a phylogenetic tree. Suppose that evolution proceeds like Brownian motion, such that changes in a trait value along a given lineage occur as small, random steps, with increases and decreases equally likely. Whenever a lineage divides, the two (or more) daughter lineages evolve independently. Under this model of Brownian motion evolution, the trait values at the tips of the phylogenetic tree follow a multivariate normal distribution with mean equal to the mean of the base of the tree, and covariance matrix $\sigma^2\mathbf{V}$ whose element $\sigma^2 v_{ij}$ for species i and j is proportional to the branch length of their shared lineage in the tree (Grafen 1989, Martins and Hansen 1997). Thus, closely related species have recent common ancestors and thus have a common evolutionary history along their shared branch length on the phylogenetic tree, leading to relatively large covariances $\sigma^2 v_{ij}$ between their trait values.

In the regression model (Eq. 2), the error terms represent variability between the observed home-range areas and the home-range areas best predicted by body size. There are many factors that could cause variation in home-range area that is not explained by body size: diet, gut physiology, leg morphology, habitat preference, social structure, etc. All of these traits likely have a phylogenetic component, with, for example, closely related species having more similar diets and gut physiologies (Garland et al. 1993). The error terms ε are the sums of the effects of all of these traits on home-range area, and due to the phylogenetic relationships among species, the covariance matrix $\sigma^2\mathbf{V}$ of the error terms should be given by the phylogenetic tree. Of course, this assumes that evolution of the unmeasured traits affecting home-range area is in fact Brownian motion; this assumption can be tested statistically, although this is beyond the scope of our present discussion (Blomberg et al. 2003).

Statistical approach

The regression problem given by Eq. 2 for the case in which all of the elements of $\mathbf{V}$ are known can be solved exactly using generalized least squares (GLS) (Garland and Ives 2000). Although there are other mathematical approaches, such as a procedure known as Independent Contrasts (Felsenstein 1985), these are just different algorithms leading to the same endpoint as GLS (Garland and Ives 2000). GLS estimates of $\mathbf{b}$ and σ^2 are

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$$
$$\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})/(n - 2) \qquad (3)$$

where $\mathbf{X}$ is the matrix whose first column is made up of ones and the second column contains the values of body size x (Judge et al. 1985: chapter 2). If there were no correlation between error terms, this would be an ordinary least squares (LS) regression, with $\mathbf{V} = \mathbf{I}$ and the estimators of $\mathbf{b}$ and σ^2 familiar from standard reference books (Neter et al. 1989). Just as in LS regression, after standardization the estimator of $\mathbf{b}$ has a t distribution, and the estimator of σ^2 has a χ^2 distribution, making it possible to calculate confidence intervals, test hypotheses, etc. (Judge et al. 1985: chapter 2).

To illustrate the GLS analysis, we simulated a single realization of Eq. 2 using the covariance matrix derived from the phylogenetic tree in Fig. 1A, and ''true'' values of b_0 and b_1 both set to zero. Treating this simulated data set as real data, we then computed GLS estimates of b_0 and b_1 using Eq. 3 (Table 2). As should be the case, the estimates of b_0 and b_1 were not statistically significantly different from zero. In contrast, if LS estimation were (incorrectly) applied to the data, the estimate of b_1 would be statistically significantly greater than zero, implying a positive relationship between home-range area and body size. Although a positive relationship is suggested by a simple plot of the data (Fig. 1B), this plot is misleading if the clusters of phylogenetically related species are ignored. The clusters show that the positive relationship is largely driven by differences among three clades. Because the species in

TABLE 2. Parameter estimates for example 1.

		GLS				LS			
Parameter	True value	Estimate	H_0: $b = 0$ $P < \|t\|$	Predicted value	95% CL	Estimate	H_0: $b = 0$ $P < \|t\|$	Predicted value	95% CL
b_0	0	−0.0436	0.5			0.5223	0.05		
b_1	0	0.2298	0.15			0.6696	0.01		
$E\{Y_h\}$				1.4413	(0.34, 2.54)			0.7045	(−0.80, 2.20)

Notes: Parameters were estimated for the generalized least squares (GLS) model and also by least squares (LS) assuming (incorrectly) that the error terms are independent. $E\{Y_h\}$ is the expected value of Y for the species marked by the arrow in Fig. 1A.

each clade are all likely to be similar in numerous ways, the similarity of their home-range areas does not necessarily implicate body size as the main determinant of home-range areas for the 15 simulated species.

Comments

Researchers are sometimes reluctant to employ methods that account for phylogenies in comparative studies because this seems to cause a ''loss of power'' in the tests they perform. In the example (Table 2), ignoring phylogeny (i.e., using LS regression) would lead to the conclusion that home-range area is positively related to body size, whereas the GLS analysis would not. Since in reality there is no relationship ($b_1 = 0$), this is not a ''loss of power'' caused by accounting for phylogeny, but instead it is GLS correctly fitting the data. In general, a statistical analysis that accounts for phylogenetic relatedness should be performed in comparative studies; assuming trait differences among species show a phylogenetic signal is a reasonable null hypothesis. If there are doubts about whether this is appropriate, diagnostics can be used to address whether there is in fact a phylogenetic signal to the data (Blomberg et al. 2003).

Although it may sometimes be the case that strong phylogenetic correlations make it difficult (i.e., require more data) to identify relationships between variables, phylogenetic correlations themselves provide information. For example, consider the problem of predicting the home-range area of a species whose body size and location on the phylogenetic tree are known; specifically, suppose that a sixteenth species identified by the arrow in Fig. 1A has body size 0.348. From this information, it is possible to calculate the predicted home-range area of this species with 95% confidence intervals using standard methods (Judge et al. 1985: chapter 2). In this example, the confidence interval has roughly the same width as that obtained (incorrectly) using LS regression (Table 2). Even though the body size of the new species is providing no information about its predicted home-range area (since $b_1 = 0$), the home-range areas of its phylogenetic relatives are providing information. Because phylogenetic correlations are structured into the GLS regression model, predictions are made using this information (see Garland and Ives 2000).

SPATIAL DATA

Example problem 2

Suppose you collected data every 10 m at 50 points along a transect through secondary forest. The data consist of soil nitrogen and aboveground net primary production (NPP) (Shaver et al. 1990). Fig. 2 shows a simulated example of a data set that might be collected in such a study, with soil nitrogen and NPP standardized so that both have mean zero and variance one. The simulation was designed so that soil nitrogen varies along the transect in an autocorrelated way, with nearby samples tending to have similar values (Fig. 2A, light line). Variation in NPP (Fig. 2A, heavy line) is positively associated with variation in soil nitrogen (Fig. 2C). In addition, there is a spatial component to variation in NPP that is not explained by soil nitrogen; plotting the difference between NPP and soil nitrogen shows spatial autocorrelation after an effect of soil nitrogen is removed (Fig. 2B). The problem is to quantify the importance of soil nitrogen on NPP and any spatial component to variation in NPP beyond that explained by soil nitrogen.

The simulated data set was generated using the general regression model given by Eq. 2, with y = NPP and x = soil nitrogen. Many forms of covariance matrices have been used to describe spatial correlation; here, we use the form (Cressie 1991):

$$\sigma^2\mathbf{V} = (1 - g)\sigma^2\begin{pmatrix} 1 & \rho & & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-1} \\ \rho^{n-1} & \rho^{n-2} & & 1 \end{pmatrix} + g\sigma^2\mathbf{I}. \quad (4)$$

This formulation splits variation in the errors into two terms. The first gives the spatial component of $\sigma^2\mathbf{V}$, with ρ measuring spatial autocorrelation. As the term is constructed, spatial correlation drops off exponentially with distance between sample locations, so that a distance of i intervals between locations leads to a correlation of ρ^i. The second term in Eq. 4 gives the nonspatial component of $\sigma^2\mathbf{V}$; the local variance of the error terms, $g\sigma^2$, is often called the ''nugget effect'' (Isaaks and Srivastava 1989). The parameter g scales the magnitude of the nugget relative to the spatial component of $\sigma^2\mathbf{V}$; the larger the value of g, the greater

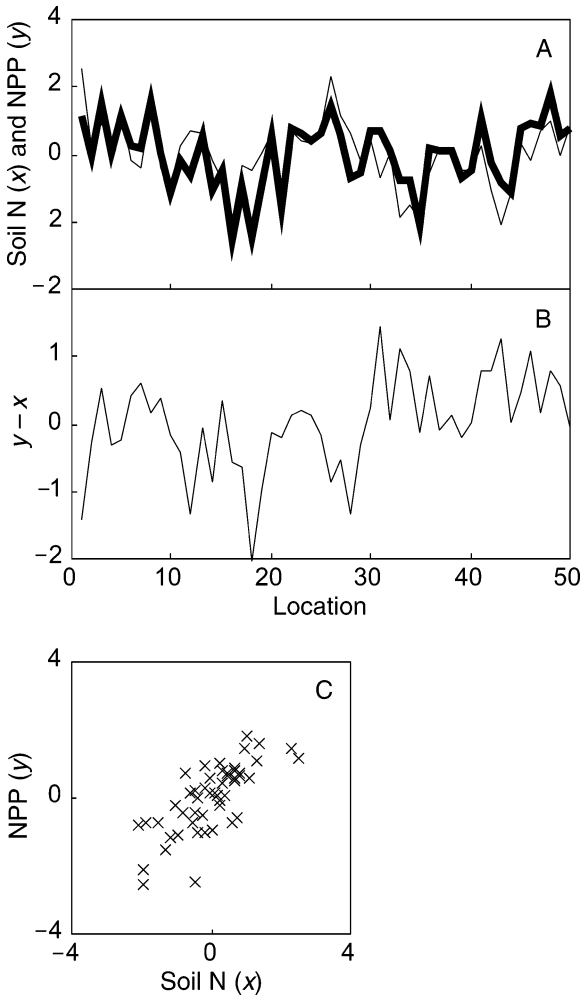

FIG. 2. (A) Soil nitrogen concentration (light line) and aboveground net primary production (heavy line) at 50 points along a transect simulated by the regression model with parameters given in Table 3. Values of soil N and NPP are standardized to have mean 0 and variance 1. (B) Difference between NPP and N at each point along the transect. (C) NPP vs. soil N.

the local variability in error terms ε relative to the spatial component.

Statistical approach

For the data set generated by simulating Eq. 2 using the covariance matrix of Eq. 4, the goal of the statistical analyses is to fit the same model and estimate the regression coefficients b_0 and b_1, and the parameters in the covariance matrix, σ^2, ρ, and g. In contrast to the example with a phylogeny (Eq. 3), the matrix $\mathbf{V}$ contains two parameters (ρ and g) that must be estimated. This makes it impossible to use GLS, for which $\mathbf{V}$ must be known. Here, we illustrate maximum-likelihood (ML) estimation.

The ML estimates for the model parameters are those values for which the data would be the most likely observation from the model (Judge et al. 1985: chapter 2, Cressie 1991). The realized values of the error terms (i.e., the residuals) are given by $(\mathbf{y} - \mathbf{Xb})$ and follow a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\mathbf{V}$. By a well-known statistical result, the log likelihood for the observed data given any set of values of b_0, b_1, ρ, g, and σ^2 is

$$l(b_0, b_1, \rho, g, \sigma^2) = -\frac{n \log 2\pi\sigma^2}{2} - \frac{\log|\mathbf{V}|}{2} - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}) \quad (5)$$

where n is the number of samples, and $|\mathbf{V}|$ is the determinant of $\mathbf{V}$ (Judge et al. 1985: chapter 2). The function $l(b_0, b_1, \rho, g, \sigma^2)$ is the log likelihood function, and the ML parameter estimates are those that maximize $l(b_0, b_1, \rho, g, \sigma^2)$.

For the simulated data in Fig. 2, the ML parameter estimates are given in Table 3. It is possible to approximate standard deviations and confidence intervals for the ML estimates. Information about the variances of the estimators can be derived from the log-likelihood function, and when the estimators are scaled by these variance terms, they follow asymptotically normal distributions (Judge et al. 1985: chapter 5). The resulting approximate standard deviations and confidence intervals approach the true standard deviations and confidence intervals only in the limit as the sample size approaches infinity, but they provide reasonable approximations for sample sizes short of infinity. Unfortunately, it is generally difficult to say how good the approximations are for a given data set without extensive simulations.

For this example, the estimated values of the parameters are fairly close to their true values, the values used to produce the simulated data. All parameters of interest other than g were identified as being statistically different from zero with 95% confidence ($\alpha = 0.05$). Importantly, both a direct effect of soil nitrogen (b_1) and a residual spatial effect of some unidentified variable(s) (ρ) were identified by the fitted model.

Comments

In the model we used for illustration, we selected a very simple description of the spatial correlation in errors ε that contained only two parameters, ρ and g. An example of a more complicated model would be

TABLE 3. Maximum-likelihood (ML) parameter estimates for example 2 with spatial correlation.

Parameter	True value	Estimate	95% CL
b_0	0	0.228	(−0.115, 0.571)
b_1	1	0.910	(0.690, 1.13)
ρ	0.8	0.695	(0.295, 1.09)
g	0.5	0.443	(0, 0.924)
σ	0.75	0.656	(0.476, 0.837)

one in which the change in correlation with distance between sample sites is unspecified, so that the correlation between samples i sites apart on the transect is the parameter ρ_i. This leads to a model with many more parameters contained in **V**. In comparing these or other models, criteria based on the log-likelihood function, such as Akaike's Information Criterion (AIC) (Akaike 1973), can be employed (Hoeting et al. 2006). These methods provide a statistical comparison among models in their ability to describe a data set when models differ in the number of parameters they contain.

Although our example confines space to one dimension along a linear transect, the same approach can be used for two- or three-dimensional space. There is considerably more mathematical bookkeeping required, however, and more computing power demanded.

Time-Series Data

Example problem 3

Suppose you have data on the population abundance of a butterfly species for 50 years (with one generation per year), and for each year you also have the abundance of its sole host plant (Fig. 3A). Host-plant abundance has decreased over the 50 years, and you want to answer two questions. Is the decrease in butterfly abundance caused by the decrease in plant abundance? And, how strong is the density dependence that affects butterfly dynamics? Although these questions do not obviously involve correlated data, the time-series nature of the data induces correlation. As shown below, the first question involves estimating regression coefficients, and the second question involves estimating the correlation structure of the data.

We simulated the butterfly and host-plant data shown in Fig. 3 using the following model:

$$x(t) - \bar{x}(t) = \rho[x(t-1) - \bar{x}(t-1)] + \alpha(t)$$
$$\bar{x}(t) = b_0 + b_1 u(t). \quad (6)$$

Here, $x(t)$ and $u(t)$ are the log abundances of butterflies and host plants, and $\alpha(t)$ are normal random variables with mean zero and variances σ^2 that are serially independent (i.e., $\alpha(t)$ and $\alpha(s)$ are independent for all $t \neq s$). The model describes changes in the difference between $x(t)$ and $\bar{x}(t)$; $\bar{x}(t)$ would be the mean (equilibrium) density of butterflies if the abundance of host plants did not change. The strength of density dependence is determined by ρ. Interpreting ρ is easiest for the situation in which $\bar{x}(t)$ does not change through time. If ρ is close to zero, $x(t)$ will remain close to $\bar{x}(t)$, giving the case of strong density dependence. In contrast, if ρ is close to one, a large (or small) value of $x(t-1) - \bar{x}(t-1)$ will likely be followed by a large (small) value of $x(t) - \bar{x}(t)$, so the butterfly population will tend to remain far from $\bar{x}(t)$. If ρ equals one, the population exhibits a random walk, with log butterfly abundance eventually reaching positive or negative infinity. Rather than assuming $\bar{x}(t)$ does not change, however, the model assumes that $\bar{x}(t)$ is a function of log host-plant abundance, $u(t)$. If, for example, b_1 were positive, then the long-term mean butterfly abundance, $\bar{x}(t)$, would be larger if the abundance of host plants was high.

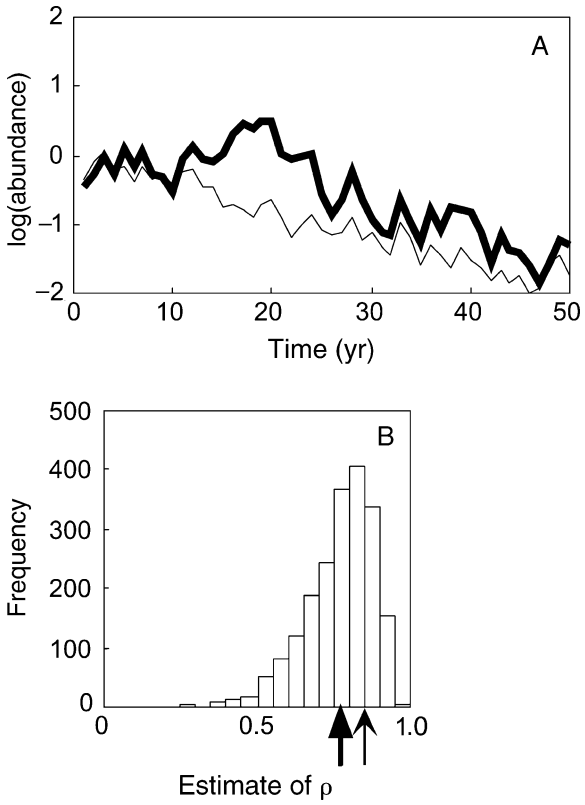

FIG. 3. Abundances of butterflies (heavy line) and their host plant (light line) for 50 butterfly generations (years) simulated from the regression model with parameters given in Table 4. (B) Bootstrap approximate distribution of the estimator of ρ for the simulated data set given in panel A. The parametric bootstrap was performed by simulating the regression model 2000 times for the parameter values estimated from the original data set. The light arrow gives the original estimate of $\rho = 0.856$ used to create the bootstrap simulated data sets, and the heavy arrow gives the mean of the approximate distribution of the estimator, $\rho = 0.770$.

Eq. 6 is an autoregressive model of order one, AR(1), with a time-varying component $u(t)$ (Harvey 1989: chapter 2, Box et al. 1994). The order of the model refers to the time lag; the model is order one, because the value of $x(t)$ depends only on the value of $x(t-1)$ and not on the value of x at more distant time lags. The equation differs from the standard formulation of regression, because values from the data on butterfly abundance, x, occur on both the left- and right-hand sides of the equation. Also, in contrast to the examples with phylogenetic and spatial correlations, the error terms $\alpha(t)$ are not correlated. Nonetheless, because val-

TABLE 4. EGLS estimates for example 3 with temporal autocorrelation.

		EGLS		ML	
Parameter	True value	Estimate	95% CL	Estimate	95% CL
b_0	0	0.303	(−0.103, 0.705)	0.303	(−0.146, 0.752)
b_1	1	0.908	(0.620, 1.19)	0.908	(0.625, 1.18)
ρ	0.9	0.856	(0.456, 0.916)	0.855	(0.816, 0.913)
σ	0.2	0.201	(0.159, 0.238)	0.195	(0.140, 0.250)

ues of $x(t)$ depend on values of $x(t-1)$, $x(t)$ and $x(t-1)$ are not independent.

In Appendix B we show that the model given by Eq. 6 can be reformulated as

$$\mathbf{x} = \mathbf{U}\mathbf{b} + \boldsymbol{\varepsilon}$$

$$\sigma^2\mathbf{V} = \sigma^2\begin{pmatrix} 1 & \rho & & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \rho^{n-1} & \rho^{n-2} & & 1 \end{pmatrix} \qquad (7)$$

where $\mathbf{U}$ is the matrix whose first column contains ones and second column contains values of $u(t)$, and $\boldsymbol{\varepsilon}$ is the vector of normally distributed error terms with mean zero and covariance matrix $\sigma^2\mathbf{V}$. This formulation has values of log butterfly abundance, $\mathbf{x}$, only on the left-hand side of the equation. Furthermore, the strength of density dependence, ρ, now governs the temporal autocorrelation between error terms; weak density dependence (ρ close to one) leads to strongly correlated errors, because the population abundance in one generation is largely determined by the abundance in the previous generation. The effect of host-plant abundance is a linear dependence of values of $x(t)$ on $u(t)$. The model given by Eq. 7 is identical to the spatial model when there is no nugget ($g = 0$).

Statistical approach

Because the time-series model has the same structure as the spatial model, it could be fit in the same way using ML estimation. To give an example of a different approach, however, we will use estimated generalized least squares (EGLS). Because the matrix $\mathbf{V}$ contains a parameter, ρ, that must be estimated, GLS cannot be used. EGLS is an extension of GLS in which components of $\mathbf{V}$ are estimated (Judge et al. 1985: chapter 5). EGLS is not commonly used in analyzing time-series data, but it nonetheless is useful to highlight some statistical points.

EGLS can be implemented as an iterative procedure. First, assume that $\rho = 0$ and calculate the GLS estimates of b_0 and b_1 using the formula given in Eq. 3. For these estimates, compute the residuals $\mathbf{r} = \mathbf{x} - \mathbf{U}\mathbf{b}$. Since the residuals $\mathbf{r}$ are realizations of the errors $\boldsymbol{\varepsilon}$, the correlation between residuals gives an estimate of the correlation between error terms, ρ. Thus, using the correlation between residuals as an estimate of ρ, estimates of b_0 and b_1 can again be calculated using the GLS formula. The residuals calculated using the new estimates of b_0 and b_1 give a new estimate of ρ, and the procedure can be repeated until estimates of b_0, b_1, and ρ converge, thus giving the EGLS estimates of b_0, b_1, and ρ, and also of σ^2 from Eq. 3.

EGLS estimators for b_0 and b_1 are unbiased, meaning that their expected values are equal to the true parameter values. However, the estimator of ρ is only asymptotically unbiased, so it may be biased unless sample sizes are very large (Judge et al. 1985: chapter 5). The standard deviation of the estimators, and hence their confidence intervals, cannot be obtained using the standard GLS formula (Eq. 3), because the standard formula assumes that the covariance matrix $\sigma^2\mathbf{V}$ is known without any uncertainty. This is not the case in the present problem due to the uncertainty in the estimate of ρ.

To obtain confidence intervals for the parameter estimates, we used parametric bootstrapping (Efron and Tibshirani 1993). Parametric bootstrapping is a simulation approach in which the statistical model fitted to the data is used to simulate a large number, say m, bootstrap data sets. For this example, we simulated Eq. 6 using the estimated values of all parameters, b_0, b_1, ρ, and σ^2, to obtain $m = 2000$ bootstrap data sets. For each of these m data sets, we estimated values of parameters, treating the bootstrap data sets as if they were real. The resulting m values of the parameters approximate the distribution of their estimators, since by definition the distribution of an estimator is the distribution of estimates that occur under the assumption that the parameter estimates and model are correct.

The EGLS estimates and their 95% confidence intervals for the data set given in Fig. 3 are displayed in Table 4. For comparison, we also computed the ML estimates and 95% confidence intervals calculated from the likelihood function. The EGLS and ML estimates are in close agreement, and are fairly close to the true parameter values used to simulate the data. Thus, in the data set the decline in host-plant abundance causes a decline in butterfly abundance ($\hat{b}_1 > 0$), and density dependence is weak ($\hat{\rho} = 0.85$), causing strong autocorrelation in butterfly abundance. The only cause for concern is the estimates of ρ, which seem low, with the true value of 0.9 towards the upper limit of the 95% confidence intervals calculated both from EGLS bootstrapping and ML approximations. Furthermore, the lower bound of the EGLS bootstrapped confidence in-

terval is considerably lower than obtained from the ML procedure.

Differences between true and estimated values suggest that parameter estimates are biased. More information can be obtained by looking at the bootstrap approximate distribution of the estimator of ρ (Fig. 3B). For the simulations used to construct the approximate distribution, the value of $\hat{\rho} = 0.856$, yet the mean of the approximate distribution is 0.770. This confirms that the EGLS estimator is biased, with the mean of the estimator less than the true mean of the process from which the data are derived. This is not a failing of the EGLS estimator alone; the ML estimator is equally biased. The explanation for the bias is suggested by the skewed distribution of the estimator. Estimates of ρ never exceed one, since values of $\rho > 1$ imply qualitatively different dynamics than those observed in the data; if $\rho > 1$, the butterfly population would diverge from $\bar{x}(t)$ exponentially. The upper limit on the value of ρ leads to skew in the estimator and also the bias. Although biased estimators are clearly not ideal, they are not uncommon. Indeed, the ML estimator of the variance of a set of data is $1/n\ \Sigma_{i=1}^{n}\ (x_i - \bar{x})^2$, whereas the unbiased estimator is the more familiar $1/(n-1)\ \Sigma_{i=1}^{n}\ (x_i - \bar{x})^2$. An advantage of bootstrapping (or simulation approaches in general) is that it reveals bias which otherwise would not be apparent.

Comments

We formulated our time-series model to reveal explicitly the correlation structure that time series introduces into data. Specifically, we formulated the model as a regression with covariance among errors given by the matrix $\sigma^2\mathbf{V}$. Most approaches, however, leave the sequence of observations in place, retaining $x(t)$ on the left-hand side and $x(t-1)$ on the right-hand side of the equation, which is more intuitively clear (Box et al. 1994).

An important issue in time-series models is how to treat the first observation, $x(1)$. In the model, all values of $x(t)$ for $t > 1$ are predicted using the previous value $x(t-1)$, but this is clearly impossible for $x(1)$. In our analyses, we assumed very little about the value of $x(0)$ occurring before the first observed value. Specifically, we assumed that $x(0)$ is selected at random from the distribution of x that occurs when $u(t)$ is zero (Appendix B). For our model, this makes $x(0)$ a normal random variable with mean zero and variance $\sigma^2/(1+\rho^2)$. This assumption gives the particularly simple form of $\mathbf{V}$ in Eq. 7. Nonetheless, this approach does not use all of the information available about $x(0)$, in particular the value of $x(1)$ and subsequent values of x(t). Using observed values of $x(t)$ to back calculate the estimate of $x(0)$ will generally lead to better parameter estimates (Box et al. 1994) and is likely to reduce bias in parameter estimates. Another approach is simply to use $x(1)$ as the first data point and ignore any information that the value of $x(1)$ might provide about the model and parameter values. This approach is called conditional least squares (CLS), because the parameters of the resulting model can be estimated using least squares regression conditional on the first value $x(1)$. Box et al. (1994) give a thorough discussion of different ways of treating the first point in the time series.

Both EGLS and ML estimators of ρ were biased, and bias is a common problem in analyses of correlated data. The best way to identify bias is through bootstrapping, either parametric bootstrapping like we used or regular bootstrapping. (In regular bootstrapping, rather than obtain values of error terms from a random number generator, the errors are selected at random, with replacement, from the residuals of the model fitted to the data; see Efron and Tibshirani 1993.) The possibility of bias is typically not stressed in statistical software packages, making some form of bootstrapping or simulation advisable. In general, simulations can be very informative about statistical analyses, often giving information about the data and analyses that is surprising.

Given that the estimators are biased, two general approaches can be used. First, bootstrapping reveals the magnitude of the bias and thereby provides a way to compensate (Efron and Tibshirani 1993). Second, the estimation procedure can be reformulated to produce estimators with less bias. This is the strategy with restricted maximum likelihood (REML), which we illustrate in the next section.

Linear Mixed Models

Example problem 4

Suppose you performed an experiment designed to investigate the factors responsible for the abundance of a predatory insect species in large agricultural fields containing its insect prey (Östman and Ives 2003). Four fields are each divided into two sections, with each section subjected to one of two treatments. One treatment is to spray an insecticide early in the growing season to reduce the initial density of prey. The second treatment is an unsprayed control. Densities of prey and predators in each of the eight field sections are sampled every week for 10 weeks, at which time the fields are harvested. During this period, prey densities increase in all sections, and although there is initially a strong treatment effect, this effect is reduced through time (Fig. 4A). Predator densities also increase with time, and predator densities in sections that were treated with insecticides tend to be lower than controls (Fig. 4B).

Several questions can be addressed with these data. First, is there a direct effect of prey density on predator abundance? The predators are highly mobile, easily capable of flying between fields, and they may aggregate in response to high prey abundances at local scales. Therefore, you suspect a large effect of prey density,

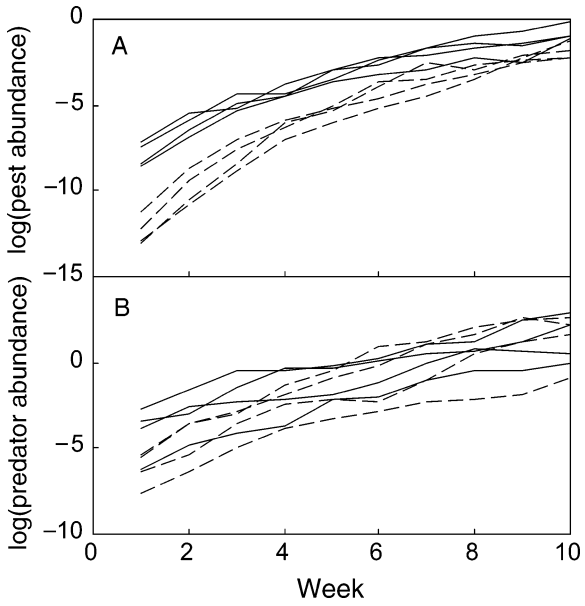

FIG. 4. Simulated experiment to investigate the factors affecting the abundance of a predator species in fields containing prey. Simulations were performed using Eq. 8 with the parameters given in Table 5. (A) Prey density in each of four fields subjected to insecticide treatment early in the season (dashed lines) or not (solid lines). (B) Corresponding densities of the predator.

which is why you designed the experiment to manipulate prey density. Second, predators are reproducing throughout the summer, so is there an overall increase in predator density through time? Third, although the insecticide treatments were designed to manipulate only prey density at the beginning of the study, is there any evidence of a residual effect of insecticides directly on the predators? Finally, for practical reasons, you could not strictly standardize the fields, so they differ in size, type of surrounding crops, etc. Thus, are there differences in predator densities among fields that are not explained by prey density, time, or treatment?

For this example, we simulated data harboring these questions using the following regression model:

$$y = b_0 + b_1x + b_2t + b_3z + u + \varepsilon$$
$$\mathrm{E}[\mathbf{uu}'] = \sigma_u^2\mathbf{G} \qquad \mathrm{E}[\boldsymbol{\varepsilon\varepsilon}'] = \sigma_\varepsilon^2\mathbf{R} \tag{8}$$

where y is the log density of predators, x is the log density of prey, t is time (week of the experiment), z is treatment (0 = control, 1 = insecticides), u indicates the field, and ε are error terms. Under the assumption that predator densities increase exponentially over the growing season, time t is treated as a continuous variable, $t = 1, 2, \ldots, 10$, rather than as a categorical variable that does not take order into account. Because predators that are in a field at one sample may remain in the field until the next sample, the density of predators is expected to be autocorrelated. Therefore, the covariance matrix for $\boldsymbol{\varepsilon}$, $\sigma_\varepsilon^2\mathbf{R}$, is constructed so that covariances between errors from different sections of fields are zero, and within any given section of a field, the correlation between $\varepsilon(t)$ and $\varepsilon(t + s)$ is ρ^s. This assumption leads to a covariance matrix for errors within sections of fields having an AR(1) structure as developed in the model for time series (Eq. 7). Finally, possible differences among fields are given by the variable u, which is assumed to be normally distributed with mean zero and variance σ_u^2. Only four values of u are selected from the normal distribution, one for each field, with larger values of σ_u^2 implying greater among-field differences.

The regression Eq. 8 is a linear mixed model (LMM), because it contains independent variables that are regarded as fixed (prey density x, time t, and treatment z) and independent variables that are themselves drawn from a random distribution, in this case the effect of field, u (Snedecor and Cochran 1989: chapter 13.9, Neter et al. 1996: chapter 24). Because u is treated as a random variable, it has a covariance matrix, which for this model is

$$\sigma_u^2\mathbf{G} = \sigma_u^2\begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{\Theta} & \boldsymbol{\Theta} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta} & \boldsymbol{\Gamma} & \boldsymbol{\Theta} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta} & \boldsymbol{\Theta} & \boldsymbol{\Gamma} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta} & \boldsymbol{\Theta} & \boldsymbol{\Theta} & \boldsymbol{\Gamma} \end{pmatrix}. \tag{9}$$

This form of $\mathbf{G}$ assumes that data are sorted first by field and then by section within fields, so that the first 20 data points correspond to the 10 points taken in each of the two sections within the same field. Matrix $\mathbf{G}$ is block diagonal, with matrix $\boldsymbol{\Gamma}$ repeated down the diagonal being a 20×20 matrix of ones, and matrix $\boldsymbol{\Theta}$ in the off-diagonal positions being a 20×20 matrix of zeros. This structure of $\mathbf{G}$ means that the values of u for samples from the same field are perfectly correlated (i.e., the same), but there is no correlation in values of u among fields. The correlation matrix $\mathbf{R}$ for error terms ε is also block diagonal, with eight 10×10 matrices along the diagonal having the AR(1) structure give in Eq. 7 and zeros in the remaining elements. Thus, within a section there is a first-order correlation of ρ. The overall covariance matrix for the model is

$$\sigma^2\mathbf{V} = \sigma_u^2\mathbf{G} + \sigma_\varepsilon^2\mathbf{R}. \tag{10}$$

To simulate the data, we first simulated values of log prey densities, x, assuming the dynamics were governed by an AR(1) process, starting at low densities to ensure a general increase in prey densities in all fields. A treatment effect on prey densities was imposed by lowering by a random amount the initial densities of prey in the sections of fields receiving insecticide treatment. Log predator densities, y, were then simulated using the model given by Eq. (8).

Statistical approach

To fit the model to the simulated data, we used restricted maximum likelihood (REML). REML is sim-

TABLE 5. REML estimates of example 4.

		Including field, u		Excluding field, u	
Parameter	True value	Estimate	H_0: $b = 0$ $P > \|t\|$	Estimate	H_0: $b = 0$ $P > \|t\|$
b_0	0	−0.233	0.8	−0.622	0.6
b_1 (prey)	0.5	0.494	0.0001	0.461	0.0001
b_2 (time)	0.2	0.244	0.0001	0.271	0.0017
b_3 (treatment)	0	0.325	0.1	0.207	0.83
σ_u^2 (field)	1	1.86			
ρ	0.8	0.574		0.957	
σ_ε^2	0.16	0.202		2.06	
AIC			98.22		111.74

ilar to ML estimation, being based on the log likelihood function, but the calculations are formulated to estimate subsets of parameters separately. For the mixed model, the variance components σ_u^2, ρ, and σ_ε^2 are estimated separately from the other parameters. To fit the model, we used PROC MIXED in SAS (Littell et al. 1996), although similar procedures exist in the S/S+ and R programming languages (Pinheiro and Bates 2000, Dalgaard 2002). REML tends to be less biased than ML, although the statistical distributions of the estimators are only known asymptotically, so statistical tests and confidence intervals are only approximate.

Fitting the model used to simulate the data shows that the analysis provides reasonable parameter estimates and correctly identifies the structure of the data (Table 5). The analysis identifies the effect of prey density and time (week of sample) on predator density, and shows no statistically significant direct effect of the insecticide treatment on predators. The autocorrelation ρ between predator densities in successive samples in the same section of a field is underestimated; the true value is 0.8 and the estimate is 0.574. As found in the time-series example, simulation studies showed that the value of ρ is consistently underestimated in the REML analysis. Specifically, parametric bootstrapping based on 1000 simulations gave a mean and 95% confidence interval 0.702 (0.449, 0.868) for the estimate of ρ when the true value was 0.8. This downward bias results from the upper bound of $\rho = 1$ required for nonexplosive predator dynamics. The estimate of the among-field variance, σ_u^2, was also biased, in this case upwards, with parametric bootstrapping giving estimated mean and confidence interval 0.0818 (0.0371, 0.1621) when the true value was 0.04. This upwards bias is caused by the lower bound of $\sigma_u^2 = 0$.

The SAS analysis does not give confidence intervals for the variance components of the model. To determine whether the overall structure of the model containing these components is warranted, we can use a model-comparison approach. Specifically, to determine whether there is reason to include among-field variability, we fit an alternative model that does not include the variable u (Table 5). The alternative model had a higher AIC value than the original, indicating a poorer fit to the data (Burnham and Anderson 1998). Interestingly, the estimate of ρ in the alternative model was much higher than in the original. This occurs because in the data there are large differences between fields, and since the alternative model is not allowed to account for this directly, it forces among-field variance into the AR(1) covariance matrix. In the alternative model, consistent differences among fields require high autocorrelation among observations within fields.

Comments

Linear mixed models give a very flexible framework for analyzing data with different types of correlations (Littell et al. 1996). Our example was a repeated-measures model, since samples were taken repeatedly from the same section within fields, and the time-series nature of the underlying biological system led us to use an AR(1) correlation structure for repeated observations within the same section. But many other types of correlation structures can be fit into the LMM format.

DISCUSSION

Correlation among samples is often unavoidable in observational data, where observations are taken, for example, from phylogenetically related species, from spatially nearby locations, or on populations through time. Even in experiments, correlation may arise, for example, from repeated measures of the same sampling unit (example 4). In experiments like the one illustrated by example 4, the researcher may want to know how a population changes through time, and this information cannot be obtained from a simple randomized experimental design that guarantees independent data. Given that correlated data are often unavoidable and sometimes advisable, ecologists should be familiar with the range of statistical approaches that have been developed for correlated data.

The four examples we have discussed in this article arose from very different contexts, yet they can be analyzed in very much the same way. The key to analyzing correlated data is to specify a statistical model that explicitly describes the correlation structure of the data. This correlation structure might be derived from phylogenies, spatial location, or dynamic processes that

dictate changes in variables through time. The correlation structure of the data is not uniquely determined by the process leading to correlation. For instance, we used the same AR(1) correlation structure to model both spatial and temporal correlation in examples 2 and 3, respectively. Furthermore, multiple sources of correlation may operate simultaneously. For instance, example 4 contains both spatial correlation (sections within the same field are more likely to be similar) and temporal correlation (data collected from the same section through time are more likely to be correlated).

Once a model specifying the correlation structure is constructed, there is usually a statistical approach for analyzing it. We illustrated a few of these approaches that can be used easily for normally distributed data: GLS, EGLS, ML, and REML. These approaches can be implemented either by writing code directly, using a matrix-friendly language such as Matlab (MathWorks 1996), or by using a statistical package such as SAS (Littell et al. 1996), S/S+ (Pinheiro and Bates 2000), or R (Dalgaard 2002, R Foundation 2004) that contains the facilities to analyze correlated data.

Ecologists are generally familiar with the problems that arise when a model does not properly fit the data, as might be revealed by a range of diagnostics, such as tests for linearity and homogeneity of variances in regression models (Neter et al. 1989). For most approaches used for correlated data, even when the model does fit the data, there is nonetheless the potential for ''incorrect'' results. We illustrated this in examples 3 and 4, in which the estimators of some parameters were biased even though the statistical model fit to the data was identical to the model used to simulate the data in the first place. Note, however, that bias in our examples was largely confined to parameters dictating the variance-covariance component of the model, ρ and σ^2, while estimates of the regression parameters b_0 and b_1 were not biased. Bias is an issue that should always be remembered when fitting all but the simplest types of statistical models, and the easiest way of checking for bias is to perform simulations.

Statistics, though sometimes seeming to complicate ecologists' lives, actually have the power to simplify ecological research by opening up more types of data and experimental designs to sound evaluation. The statistical approaches we described in this article are not difficult, and their flexibility makes them frequently valuable for ecological research.

Acknowledgments

The participants of the workshop on ''Evaluating evidence in ecological data: alternatives to statistical hypothesis testing'' provided much of the stimulus for this article. We thank B. J. Cardinale, J. Forester, T. Garland, K. J. Tilmon, M. G. Turner, and the Turner lab for suggestions on the manuscript. This work was funded in part by grants DEB-0196384 and DEB-9806953 from the U.S. National Science Foundation.

Literature Cited

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akadeiai Kiado, Budapest, Hungary.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution **57**:171–745.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. Time series analysis: forecasting and control. Prentice Hall, Englewood Cliffs, New Jersey, USA.

Burnham, K. T., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer, New York, New York, USA.

Cressie, N. A. C. 1991. Statistics for spatial data. John Wiley and Sons, New York, New York, USA.

Dalgaard, P. 2002. Introductory statistics with R. Springer-Verlag, New York, New York, USA.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York, New York, USA.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist **125**:1–15.

Garland, T., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. Systematic Biology **42**:265–292.

Garland, T., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology **41**:18–32.

Garland, T., and A. R. Ives. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. American Naturalist **155**:346–364.

Grafen, A. 1989. The phylogenetic regression. Transactions of the Royal Society of London B, Biological Sciences **326**: 119–157.

Harvey, A. C. 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge, UK.

Hoeting, J. A., R. A. Davis, A. A. Merton, and S. E. Thompson. 2006. Model selection for geostatistical models. Ecological Applications **16**:87–98.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**: 197.

Isaaks, E. H., and R. M. Srivastava. 1989. Applied geostatistics. Oxford University Press, New York, New York, USA.

Ives, A. R., B. Dennis, K. L. Cottingham, and S. R. Carpenter. 2003. Estimating community stability and ecological interactions from time-series data. Ecological Monographs **73**:301–330.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T.-C. Lee. 1985. The theory and practice of econometrics. John Wiley and Sons, New York, New York, USA.

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. SAS system for mixed models. SAS Institute, Inc., Cary, North Carolina, USA.

Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. American Naturalist **149**:646–667. Erratum **153**:448.

MathWorks, I. 1996. MATLAB. MathWorks, Inc., Natick, Massachusetts, USA.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. McGraw-Hill, New York, New York, USA.

Neter, J., W. Wasserman, and M. H. Kutner. 1989. Applied linear regression models. Richard D. Irwin, Inc., Homewood, Illinois, USA.

Östman, Ö., and A. R. Ives. 2003. Scale-dependent indirect interactions between two prey species through a shared predator. Oikos **102**:505–514.
Pinheiro, J. C., and D. M. Bates. 2000. Mixed-effects models in S and S-PLUS. Springer-Verlag, New York, New York, USA.
R Foundation. 2004. R: an open source implementation of the S language. ⟨http://www.r-project.org/⟩
Shaver, G. R., K. J. Nadelhoffer, and A. E. Giblin. 1990. Biogeochemical diversity and element transport in a heterogeneous landscape, the North Slope of Alaska. Pages 105–126 *in* M. G. Turner and R. H. Gardner, editors. Quantitative methods in landscape ecology. Springer-Verlag, New York, New York, USA.
Snedecor, G. W., and W. G. Cochran. 1989. Statistical methods. Iowa State University Press, Ames, Iowa, USA.

APPENDIX A

This appendix gives a brief list of statistical references to the methods we illustrated.

Introductory texts

Larsen, R. J., and M. L. Marx. 1981. An introduction to mathematical statistics and its applications. Prentice-Hall, Englewood Cliffs, New Jersey, USA.
Neter, J., W. Wasserman, and M. H. Kutner. 1989. Applied linear regression models. Richard D. Irwin, Homewood, Illinois, USA.

Advanced texts

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York, New York, USA.
Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T.-C. Lee. 1985. The theory and practice of econometrics. John Wiley and Sons, New York, New York, USA.
Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. McGraw-Hill, New York, New York, USA.
Snedecor, G. W., and W. G. Cochran. 1989. Statistical methods. Iowa State University Press, Ames, Iowa, USA.

Spatial models

Cressie, N. A. C. 1991. Statistics for spatial data. John Wiley and Sons, New York, New York, USA.
Goovaerts, P. 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York, New York, USA.
Haining, R. 1990. Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge, UK.
Isaaks, E. H., and R. M. Srivastava. 1989. Applied geostatistics. Oxford University Press, New York, New York, USA.
Upton, G. J. G., and B. Fingleton. 1985. Spatial data analysis by example. John Wiley and Sons, New York, New York, USA.

Time-series models

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. Time series analysis: forecasting and control. Prentice Hall, Englewood Cliffs, New Jersey, USA.
Fuller, W. A. 1996. Introduction to statistical time series. John Wiley and Sons, New York, New York, USA.
Harvey, A. C. 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge, UK.
Harvey, A. C. 1993. Time series models. Harvester Wheatsheaf, New York, New York, USA.
Reinsel, G. C. 1997. Elements of multivariate time series analysis. Springer-Verlag, New York, New York, USA.

Computational methods/programs

Spatial models.—

Kaluzny, S. P., S. C. Vega, T. P. Cardoso, and A. A. Shelly. 1998. S+ SpatialStats: user's manual for Windows and UNIX. Springer-Verlag, New York, New York, USA.

Time-series models.—

Brocklebank, J. C., and D. A. Dickey. 2003. SAS for forecasting time series. John Wiley, Hoboken, New Jersey, USA.
Venables, W. N., and B. D. Ripley. 2002. Modern applied statistics with S. Springer-Verlag, New York, New York, USA.

Mixed models.—

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. SAS system for mixed models. SAS Institute, Cary, North Carolina, USA.
Pinheiro, J. C., and D. M. Bates. 2000. Mixed-effects models in S and S-PLUS. Springer-Verlag, New York, New York, USA.

APPENDIX B

This appendix reformulates the time-series model given by Eq. 6 as a generalized regression model given by Eq. 7.

The time-series model can be written in matrix form as

$$\mathbf{x} - \bar{\mathbf{x}} = \rho(\mathbf{x}_B - \bar{\mathbf{x}}_B) + \boldsymbol{\alpha} \tag{B1}$$

where $\mathbf{x}$ and $\bar{\mathbf{x}}$ are the vectors of values of $x(t)$ and $\bar{x}(t)$, and $\mathbf{x}_B$ and $\bar{\mathbf{x}}_B$ contain zeros as their first elements and values of $x(t)$ and $\bar{x}(t)$ for $t = 1$ to $n - 1$ as their remaining elements. Vector $\boldsymbol{\alpha}$ contains $x(0)$ as its first element and values of $\alpha(t)$ as its remaining elements. This formulation, with the first element of $\mathbf{x}_B$ being zero and the first element of $\boldsymbol{\alpha}$ being $x(0)$, is necessitated because, although there are n observed values of x, there are only $n - 1$ predicted values of x (and $n - 1$ values of α), since the first predicted value of x is $x(2)$. In order to have $\mathbf{x}$ contain all of the observed data, we include an unobserved value $x(0)$, but since it is unobserved, we treat it as a random variable and hence place it in $\boldsymbol{\alpha}$.

Eq. B1 can be simplified using the backwards operator

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & & 0 \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & & 0 \end{pmatrix} \tag{B2}$$

which maps values of $\mathbf{x}$ onto values of $\mathbf{x}_B$: $\mathbf{x}_B = \mathbf{B}\mathbf{x}$. From Eq. B1,

$$\begin{aligned}\mathbf{x} - \bar{\mathbf{x}} &= \rho\mathbf{B}(\mathbf{x} - \bar{\mathbf{x}}) + \boldsymbol{\alpha}\\ (\mathbf{I} - \rho\mathbf{B})(\mathbf{x} - \bar{\mathbf{x}}) &= \boldsymbol{\alpha}\\ \mathbf{x} &= \bar{\mathbf{x}} + (\mathbf{I} - \rho\mathbf{B})^{-1}\boldsymbol{\alpha}\\ \mathbf{x} &= \bar{\mathbf{x}} + \boldsymbol{\varepsilon}. \qquad \text{(B3)}\end{aligned}$$

Because $\boldsymbol{\varepsilon}$ is a linear transformation of $\boldsymbol{\alpha}$, $\boldsymbol{\varepsilon}$ is normally distributed with mean zero and covariance matrix

$$\begin{aligned}\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') &= \mathrm{E}\{(\mathbf{I} - \rho\mathbf{B})^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}'[(\mathbf{I} - \rho\mathbf{B})^{-1}]'\}\\ &= (\mathbf{I} - \rho\mathbf{B})^{-1}\boldsymbol{\Omega}[(\mathbf{I} - \rho\mathbf{B})^{-1}]' = \sigma^2\mathbf{V} \qquad \text{(B4)}\end{aligned}$$

where $\boldsymbol{\Omega}$ is the covariance matrix of $\boldsymbol{\alpha}$. Since the values of $\boldsymbol{\alpha}$ are assumed to be independent, $\boldsymbol{\Omega}$ is a diagonal matrix with diagonal elements σ^2 except for the first diagonal element. The first diagonal element corresponds to $x(0)$, the unobserved value of x before the first observed value. If we assume that nothing is known about $x(0)$ other than that it is pulled from the stationary distribution of $x(t)$ that would occur in the absence of changes in host-plant abundance, $u(t)$, then the variance in $x(0)$ is $\sigma^2/(1 + \rho^2)$ (Ives et al. 2003). With this formulation of $\boldsymbol{\Omega}$, $\mathbf{V}$ is given by Eq. 7.

SUPPLEMENT

Annotated computer code in Matlab and SAS for performing the simulations and analyses in examples 1–4 (*Ecological Archives* A016-002-S1).