

Models for Ecological Time Series Steve Carpenter, Zoology 535

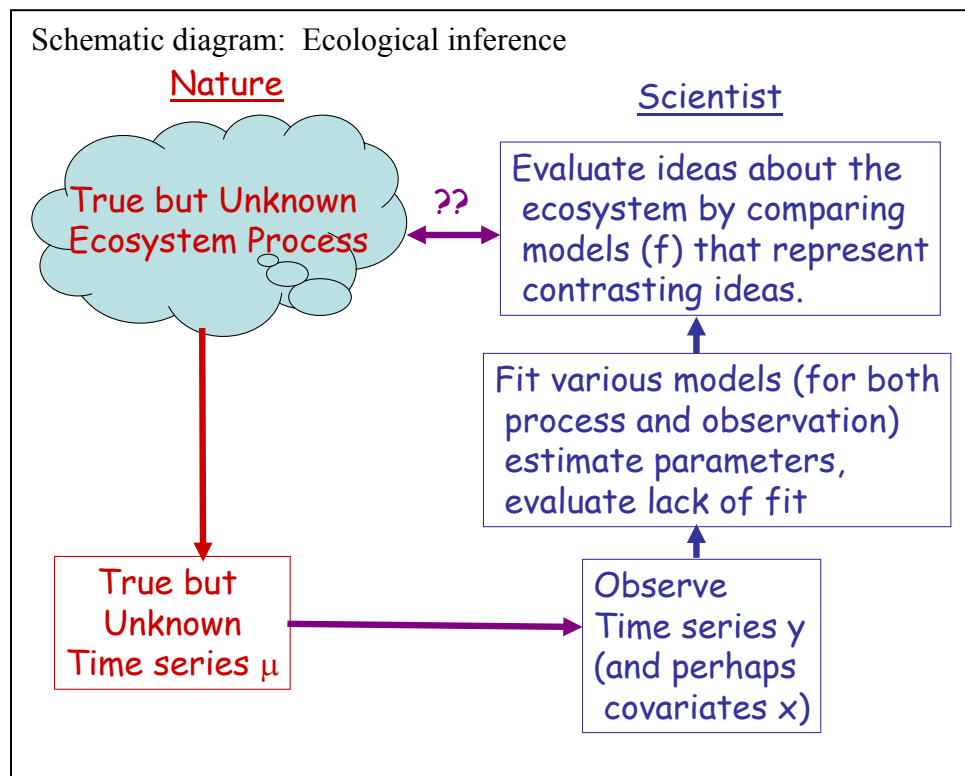
Models for changes in ecological variables over time require some additional analytical tools, beyond those we have used so far. Generally, such models relate values at a particular time to values at a previous time. For example, many ecological models have the general form

$$y_t = f(y_{t-1}, x_t, \theta) + \varepsilon_t \quad [1]$$

where y is an ecological time series, x represents covariates, θ represents parameters, and ε is a time series of residuals or errors. The time series can be a vector, and there can be multiple parameters. Covariates are optional, and there can be any number of covariates. The function f can be linear or nonlinear.

Equation 1 is called a process equation, or transition equation, and the residuals ε are called "process errors". They represent errors due to using the wrong process (f) to explain how nature generated the observations.

Statistical time series analysis emphasizes the removal of autocorrelation from the time series y . This can be important for ecological time series. However, most ecological time series are too short for detailed analysis of autocorrelation. Also, the ecological research question is usually "what model for the ecosystem (represented by f) seems most consistent with the data"? Thus ecologists usually want to examine several models and compare their fit to the data using some statistic such as AIC. This analysis may or may not involve removal of autocorrelation, depending on the details of the particular case.



Observation error, or measurement error in the y's, must often be considered when modeling ecological time series (see figure on preceding page). This calls for another equation, the observation equation, which relates y to a true but unknown value μ and observation errors v :

$$y_t = \mu_t + v_t \quad [2]$$

Often it is possible to estimate μ and v using replicate samples at each point in time.

There are many ways to address the problem of fitting equations 1 and 2 to data. We will cover three of the most common options: process error estimation, observation error estimation, and Kalman filter estimation.

Process Error Estimation

If measurements show that observation errors are small, it may be OK to ignore them and fit equation 1 as if observation errors can be ignored. Sometimes this procedure is called "pure process error fitting" because it assumes that all the error (or almost all of it) is due to mis-specification of the true ecological process. The pseudocode is as follows:

1. Using initial guesses of the parameters θ , calculate each estimated value of y using the previous observed value of y:

$$\hat{y}_t = f(y_{t-1}, x_t, \theta)$$

2. Calculate process error

$$\varepsilon_t = y_t - \hat{y}_t$$

3. Calculate the lack-of-fit statistic (sum of squared errors, negative log likelihood, etc.) and find the parameters that minimize it, using an iterative optimization program on the computer (e.g. the function 'nlm' in R, or 'fminsearch' in Matlab).

Observation Error Estimation

If observation errors are thought to be large, it may be best to assume that all the error is observation error. This procedure is sometimes called "pure observation error fitting" because it proceeds as if the true ecological process is known, but the observations are imperfect. The pseudocode is as follows:

1. Set the initial estimated value of y equal to the first observation:

$$\hat{y}_1 = y_1$$

2. Using initial guesses of the parameters θ , calculate each estimated value of y using the previous estimated value of y (*not* the previous observed value of y as in process-error fitting):

$$\hat{y}_t = f(\hat{y}_{t-1}, x_t, \theta)$$

3. Calculate errors

$$\varepsilon_t = y_t - \hat{y}_t$$

4. Calculate the lack-of-fit statistic (sum of squared errors, negative log likelihood, etc.) and find the parameters that minimize it, using an iterative optimization program on the computer (e.g. the function 'nlm' in R, or 'fminsearch' in Matlab).

Kalman Filter Estimation

If the observation errors can be estimated (e.g. from replicates), it is possible to account for both observation and process error using an algorithm called the Kalman Filter.

If the function f is nonlinear in parameters (i.e. $df/d\theta$ is a function of one or more parameters), then the usual Kalman Filter will not work. There is an alternative called the Extended Kalman Filter (EKF). However, most ecological time series are not long enough to yield good results with the EKF.

If the function f is linear in parameters (i.e. $df/d\theta$ is not a function of any parameter) then the usual Kalman Filter often performs well.

First rewrite the process equation as

$$\mathbf{\alpha}_t = \mathbf{B} \mathbf{\alpha}_{t-1} + \mathbf{C} \mathbf{u}_t + \varepsilon_t$$

where bolding indicates quantities that can be vectors or matrices. Here $\mathbf{\alpha}$ is the true but unknown value of y , \mathbf{B} is the parameters that connect y over time, \mathbf{u} is the covariates x , and \mathbf{C} is the parameters for the effects of the covariates. As before, ε represents process error, which is distributed multivariate normal with mean 0 and covariance matrix \mathbf{Q} .

$$\varepsilon \sim N(0, \mathbf{Q})$$

The observation equation is

$$\mathbf{y}_t = \mathbf{\alpha}_t + \mathbf{v}_t, \quad \mathbf{v} \sim N(0, \mathbf{H})$$

The problem solved by the Kalman Filter is this: given measurements of \mathbf{y} and the observation covariance matrix \mathbf{H} , estimate \mathbf{B} , \mathbf{C} and \mathbf{Q} .

The pseudocode is as follows.

1. Guess initial values for \mathbf{B} , \mathbf{C} and \mathbf{Q} .

2. Set the initial estimator of α , \mathbf{a}_1 , equal to the first observed \mathbf{y} : $\mathbf{a}_1 = \mathbf{y}_1$.
3. Set the process covariance for the first time step, \mathbf{P}_1 , equal to \mathbf{Q} .
4. Starting with time step 2, build time series of \mathbf{a} and \mathbf{P} by iterating through the following sequence of calculations for each time step:

4a. Predict next \mathbf{a} and \mathbf{P} based on information in preceding time steps:

$$\mathbf{a}_{t|t-1} = \mathbf{B} \mathbf{a}_{t-1} + \mathbf{C} \mathbf{u}_t$$

$$\mathbf{P}_{t|t-1} = \mathbf{B} \mathbf{P}_{t-1} \mathbf{B}' + \mathbf{Q}$$

4b. Update the predictions by correcting for the information in the new observation made at the time step:

$$\mathbf{v}_t = \mathbf{y}_t - \mathbf{a}_{t|t-1}$$

$$\mathbf{F}_t = \mathbf{P}_{t|t-1} + \mathbf{H}_t$$

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} + \mathbf{P}_{t|t-1} \mathbf{F}_t^{-1} \mathbf{v}_t$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1} \mathbf{F}_t^{-1} \mathbf{P}_{t|t-1}$$

5. For a single time step, the negative log likelihood is

$$-\log(L_t) = 0.5 N \log(2\pi) + 0.5 \log(\det(\mathbf{F}_t)) + 0.5 \mathbf{v}_t' \mathbf{F}_t^{-1} \mathbf{v}_t$$

These terms are summed up over all the time steps to calculate the overall negative log likelihood.

6. Find the values of \mathbf{B} , \mathbf{C} and \mathbf{Q} that minimize the negative log likelihood, using an iterative optimization program on the computer (e.g. the function 'nlm' in R, or 'fminsearch' in Matlab).

References

Harvey, A.C., 1989, *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge University Press.

Hilborn, R. and M. Mangel. 1997. *The Ecological Detective – Confronting Models with Data*. Princeton Univ. Press, N.J.