ISRAEL JOURNAL OF ECOLOGY & EVOLUTION, Vol. 52, 2006, pp. 1–10

ON FITTING A MODEL TO A POPULATION TIME SERIES WITH MISSING VALUES

OREN BARNEA,[a,*] ANDREW R. SOLOW,[b] AND LEWI STONE[a]
[a]*Biomathematics Unit, Department of Zoology, Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel*
[b]*Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA*

ABSTRACT

Existing methods for fitting a population model to time series data typically assume that the time series is complete. When there are missing values, it is common practice to substitute interpolated values. When the proportion of values that are missing is large, this can lead to bias in model-fitting. Here, we describe a maximum likelihood approach that allows explicitly for missing values. The approach is applied to a long weekly time series of the dinoflagellate *Peridinium gatunense* in Lake Kinneret, Israel, in which around 35% of the values are missing.

Keywords: ecological time series, maximum likelihood, missing values, Lake Kinneret, phytoplankton

INTRODUCTION

One of the basic goals of theoretical ecology research is finding the factors controlling and maintaining the size of a population (May, 1974; Den Boer and Reddingius, 1996). A very obvious controlling factor is the density of the population, and the stabilizing effect of density-dependence—when a population is relatively small it has a tendency to grow quickly, and when a population is large its growth is inhibited. However, it is not always a straightforward matter to find evidence for density dependence in real-world data sets, and success often hinges on the quality of the data and the length and regularity of sampling. As a particular example, Fig. 1 shows a time series of the density of the phytoplankton *Peridinum gatunense* in Lake Kinneret, Israel over the period 1970-1994. The population dynamics of this dinoflagellate plays an important role in the Kinneret ecosystem (Berman and Pollingher, 1974; Berman et al., 1995). One way to understand the dynamics of a population such as this one is by fitting a model to a time series of population size or density.

Fitting a model of population dynamics to time series data is greatly facilitated when the observations are regularly spaced in time. When an otherwise regularly spaced time series has a small number of missing values, it is common practice to fill in the missing

*Author to whom correspondence should be addressed. E-mail: oren.barnea@gmail.com
Accepted April 2006.

values by interpolation and to analyze the resulting time series as if it were complete (Steven and Glombitza, 1972; Jassby and Powell, 1990; Solow et al., 2003). Common interpolation methods include standard linear interpolation or more complicated cubic or higher order splines. Although the statistical properties of the interpolated values—which are the result of filtering the original observations—are different from those of the true time series, the effect on the analysis is likely to be small provided the missing observations are few. This is not the case for the time series shown in Fig. 1. As described in more detail below, this time series has a basic sampling interval of approximately 1 week with a substantial number of gaps of different duration. The purpose of this paper is to describe an approach to modeling this time series that accounts in an explicit way for the missing values. Related work focusing on linear time series models includes that of Kohn and Ansley (1986). However, since the model considered here is nonlinear, the methods described in Kohn and Ansley (1986) are not applicable.

The remainder of the paper is organized in the following way. Some background on the ecological system of Lake Kinneret is given in the following section. The basic statistical model and methods are described in the third section, followed in the fourth section by an application of these methods to the time series shown in Fig. 1. The final section contains some concluding remarks.

LAKE KINNERET PHYTOPLANKTON BLOOM DYNAMICS

The bloom of *Peridinium* occurs in spring (usually February) in most years in Lake

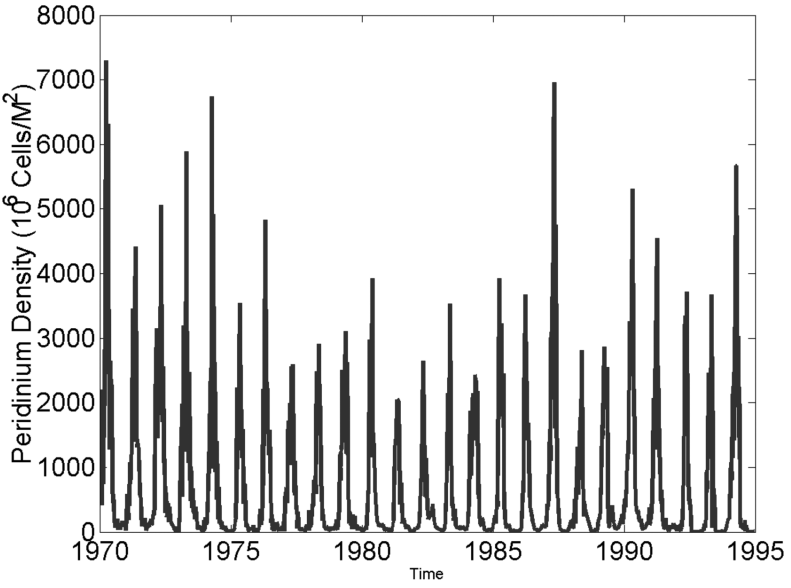

Fig. 1. *Peridinium* density (10^6 cells/M^2) between January 1970 and December 1994.

Kinneret. Blooms are periods of rapid explosive growth in population size, followed by rapid decrease. Between two blooms the population is almost stationary. Pollingher and Serruya (1976) documented the exponential growth of the bloom in the growth phase by fitting the equation

$$X_t = X_0 e^{kt}$$

where X_t is the number of *Peridinium* cells per milliliter at time *t*. This allowed them to estimate cell doubling times of between 17–40 days for the years 1969–1974. During the bloom almost the entire lake surface appears brown or coffee-colored. *Peridinium* dominates to the extent that it represents some 95% of the total algal biomass in the lake and reaches concentrations of several thousand cells per milliliter.

The conditions that terminate the bloom still remain uncertain. The bloom itself is affected little by grazing since *Peridinium* cells are not eaten by most fish species or any of the zooplankton species. More likely, with the passage of spring, the waters warm up, nutrients become depleted, CO_2 limitation sets in, and physical conditions become less favorable. At the end of the bloom, the *Peridinium* cells form cysts and sink to the sediment where they lie until resuspension the next year. Figure 2 provides a simplified visual summary of the main ecosystem features relating to the dynamics in Lake Kinneret over a typical year. Initially, with winter overturn when stratification of the lake

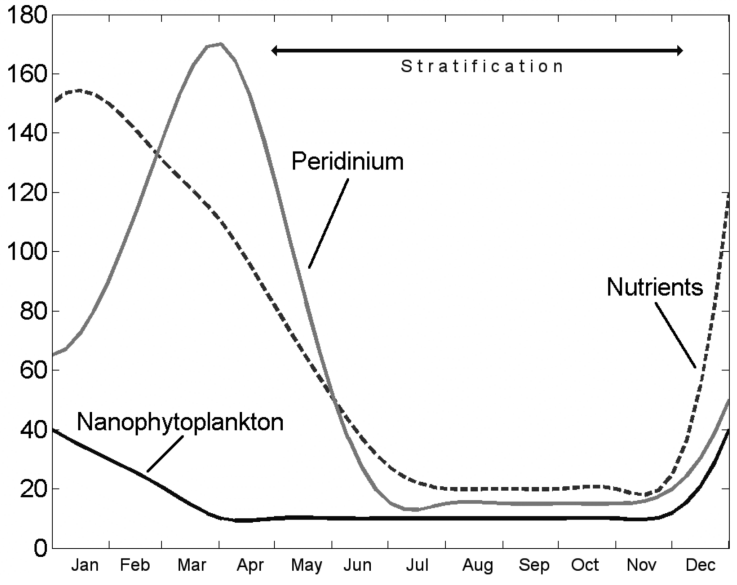

Fig. 2. Schematic summary of *Peridinium*, nutrients, and nanophytoplankton dynamics in Lake Kinneret in a typical year.

breaks down, there is an injection of nutrient supplies into the water column, *Peridinium* cysts are resuspended and bloom initiation begins soon after. As nutrient levels fall, the bloom collapses. By this stage thermal stratification of the lake is well underway. Over summer, phytoplankton biomass (mainly nanophytoplankton) remains constant but at relatively low levels.

Although undocumented, it is believed that *Peridinium* blooms have occurred for hundreds of years in Lake Kinneret. However, in 1996 the *Peridinium* bloom failed to appear and since then has lost its well-documented regularity (Berman et al., 1995). In recent years variation in bloom intensity and timing is much greater than it was between 1970 and 1995. This is likely to be an indication of major change in the Lake Kinneret ecosystem, but the precise cause remains unknown.

MODEL AND METHODS

Let X_t be the density of a population in period t. Although the model that we fit in the next section is slightly more complicated, assume for now that the dynamics of this population are governed by the first-order nonlinear autoregressive model:

$$X_t = f_\theta(X_{t-1}) \exp(\varepsilon_t) \tag{1}$$

where f_θ is a nonnegative nonlinear function that depends on the parameter θ (which is typically vector-valued) and ε_t is a normal process error with mean 0 and variance σ^2. It is convenient to rewrite (1) as

$$\log X_t = g_\theta(X_{t-1}) + \varepsilon_t \tag{2}$$

where $g_\theta = \log f_\theta$. Both θ and σ^2 are unknown and interest centers on statistical inference about the former based on a time series of observations of density. We will focus on methods based on the likelihood. The likelihood is defined as the joint probability density of the observations viewed as a function of the unknown parameters.

To begin with, suppose that the time series consists of regularly spaced observations $x_1, x_2, ..., x_n$ of $X_1, X_2, ..., X_n$. The Markovian property of the model in (1) ensures that, conditional on $X_1 = x_1$, the likelihood factors as

$$L(\theta, \sigma^2 \mid x_1) = \prod_{t=2}^{n} p(x_t \mid x_{t-1}) \tag{3}$$

where $p(x_t \mid x_{t-1})$ is the conditional probability density of x_t given $X_{t-1} = x_{t-1}$. For economy, the dependence of this probability density on the unknown parameters θ and σ^2 is suppressed in the notation. Under the model in (1), the conditional distribution of X_t given that $X_{t-1} = x_{t-1}$ is lognormal with parameters $g_\theta(x_{t-1})$ and σ^2, so that

$$p(x_t \mid x_{t-1}) = \frac{1}{x_t \sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (\log x_t - g_\theta(x_{t-1}))^2 \right) \tag{4}$$

It follows that, apart from factors not involving the unknown parameters, the log likelihood is

$$\log L(\theta,\sigma^2 \mid x_1) = -\frac{n-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(\log x_t - g_\theta(x_{t-1}))^2 \tag{5}$$

So, for example, the maximum likelihood (ML) estimate of θ is given by the ordinary least squares fit of $g_\theta(x_{t-1})$ to $\log x_t$. Unless g_θ is linear, this fitting has to be performed numerically.

Suppose now that there is a single missing observation in period m. One option is to use the available observations to reconstruct the missing value—under the simplest such scheme, the reconstructed value is $\hat{x}_m = (x_{m-1} + x_{m+1})/2$ —and to proceed as if the time series were complete. As noted, this option is reasonable in the case of a small number of missing observations, but it can cause problems otherwise. A second option is to condition on both x_1 and x_{m+1} and to split the conditional likelihood as

$$L(\theta,\sigma^2 \mid x_1, x_{m+1}) = \prod_{t=2}^{m-1} p(x_t \mid x_{t-1}) \prod_{t=m+2}^{n} p(x_t \mid x_{t-1}) \tag{6}$$

This option is also reasonable when the number of missing observations is small. Otherwise, it can involve conditioning on a large number of observations, effectively reducing the sample size for inference. A third option is to write the conditional likelihood given $X_1 = x_1$ as

$$L(\theta,\sigma^2 \mid x_1) = \prod_{t=2}^{m-1} p(x_t \mid x_{t-1})\, p(x_{m+1} \mid x_{m-1}) \prod_{t=m+2}^{n} p(x_t \mid x_{t-1}) \tag{7}$$

where

$$p(x_{m+1} \mid x_{m-1}) = \int_0^\infty p(x_{m+1} \mid x_m)\, p(x_m \mid x_{m-1})\, dx_m \tag{8}$$

and where the terms in the integral have the form given in (4). Although this integral cannot be expressed in closed form, it is straightforward to evaluate it numerically and therefore to evaluate the complete likelihood conditioning only on x_1. In the next section, we use a combination of the second and third options to fit a model to the time series shown in Fig. 1.

APPLICATION

Let X_1 be the phytoplankton biomass in week t. We will adopt the model:

$$X_t = X_{t-1} \exp(\beta + S_t + \gamma X_{t-1}) \exp(\varepsilon_t) \tag{9}$$

where S_t is a periodic function with the period of 1 year representing the effect on phytoplankton biomass of regular seasonal cycles in environmental conditions (e.g., temperature, light, nutrients). Although it would be preferable to capture these effects directly by including the appropriate environmental variables in the model, the required information is not available. The seasonal effect will be modeled as

$$S_t = \sum_{j=1}^{k} A_j \sin(\omega\, j\, d_t) + B_j \cos(\omega\, j\, d_t) \tag{10}$$

where $\omega = 2\pi / 365$ and d_t is the date of observation X_t measured in days from the beginning of the observation period. The number of harmonics k and the amplitudes A_j and B_j ($j = 1, 2, \ldots, k$) are unknown and must be estimated. The unknown parameter γ reflects the strength of any density dependence in the phytoplankton dynamics. Particular interest centers on testing the null hypothesis $H_0 : \gamma = 0$ of density independence against the one-sided alternative hypothesis $H_0 : \gamma < 0$ of density dependence. In summary, under this model, the conditional distribution of X_t given $X_{t-1} = x_{t-1}$ is lognormal with mean

$$\mu_t = \log x_{t-1} + \beta + S_t + \gamma\, x_{t-1} \tag{11}$$

and variance σ^2.

In the notation of the previous section, the parameter θ of eq 1 includes the scale factor β, the number of harmonics k in the seasonal component, and, for each harmonic, the amplitudes A_j and B_j , and the density dependence parameter γ.

The time series in Fig. 1 to which the model described above was fitted consists of a total of 834 observations covering a period of 9017 days or approximately 1288 weeks (1970–1995). The basic sampling interval in this time series is approximately weekly, although there is a large number of gaps (i.e., around 35% of the weekly observations are missing). The time series can be divided into 70 segments according to the criterion that, within each of these segments, the intervals between successive observations are either approximately 1 week or, when a weekly observation is missing, approximately 2 weeks. The interval between the last observation in each segment and the first observation in the succeeding segment is greater than 2 weeks. The average number of observations per segment is around 12 and the average number of missing weekly observations per segment is around 3.

We fit the model in (1) to the time series in Fig. 1 by conditioning on the initial observation in each of the 70 segments described above. Missing weekly observations within each segment were treated explicitly in the likelihood via (8). The number of harmonics k in the seasonal component was estimated by increasing its value from 1 until the additional pair of terms was found to be non-significant at the 0.05 level by the likelihood ratio (LR) test. The value of k selected in this way is 2. The maximum likelihood (ML) estimates of the remaining parameters are

$$\hat{\beta} = 0.204$$

$$\hat{A}_1 = 0.367 \quad \hat{B}_1 = 0.066 \quad \hat{A}_2 = -0.074 \quad \hat{B}_2 = -0.177$$

$$\hat{\gamma} = -0.0003$$

$$\hat{\sigma} = 0.492$$

The maximized value of the log likelihood is around –990. A plot of $\log X_t$ is given in Fig. 3a along with the fitted values from the model in Fig. 3b. These fitted values are the one-step-ahead predictions given by:

$$\hat{X}_{t+1} = X_t \exp(\hat{\beta} + \hat{s}_t + \hat{\gamma} X_t + \frac{\hat{\sigma}^2}{2}) \tag{12}$$

As noted, particular interest centers on whether the variations in phytoplankton biomass exhibit density dependence. This issue can be addressed in the context of the model in (9) by testing the null hypothesis $H_0 : \gamma = 0$ against the one-sided alternative hypothesis $H_0 : \gamma < 0$. To do so, it is again natural to use the likelihood ratio test. The LR test statistic is twice the difference between the value of the log likelihood maximized under H_1 and H_0, respectively. In this case, under H_0, the LR statistic has an approximate chi-square distribution with 1 degree of freedom. To calculate the LR statistic, it is necessary to fit the model by maximum likelihood under H_0—that is, under the constraint that $\gamma = 0$. The ML estimates of the remaining parameters are:

$$\hat{\beta} = 0.09$$
$$\hat{A}_1 = 0.121 \quad \hat{B}_1 = 0.194 \quad \hat{A}_2 = 0.062 \quad \hat{B}_2 = -0.115$$
$$\hat{\sigma} = 0.534$$

The corresponding value of the log likelihood is around –1055, so the LR statistic is LR = 2(–990 + 1055) = 130. This is extremely significant and the null hypothesis of density independence can be decisively rejected.

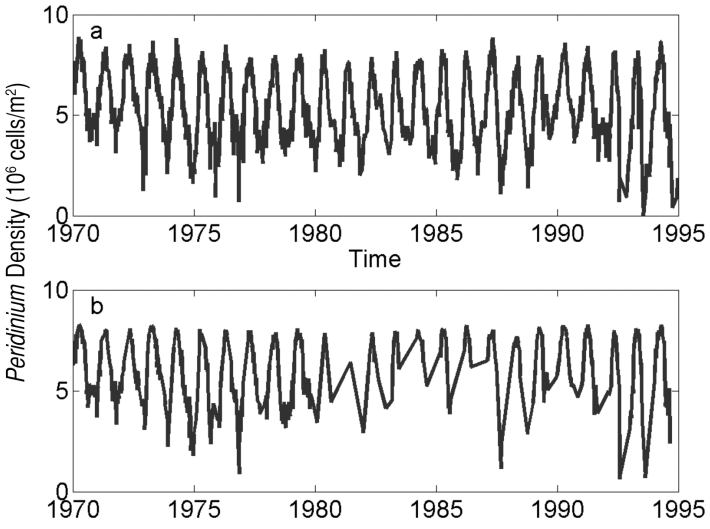

Fig. 3. (a) Log *Peridinium* density, January 1970–December 1994. (b) One-week-ahead predictions of Log *Peridinium* density, January 1970–December 1994. Note that visually, some smoothness is lost in the graph of predictions due to segmentation of the data.

As a graphical check on the adequacy of the fitted model, in Fig. 4, the partial residuals

$$R_t = \log(X_t / X_{t-1}) - (\hat{\beta} + \hat{S}_t) \tag{13}$$

are plotted against X_{t-1} for all cases in which successive weekly observations are available. Here, $\hat{S}_t$ denotes the fitted seasonal component. Note that according to (9)

$$R_t = \gamma X_{t-1} + \varepsilon_t \tag{14}$$

and Fig. 4 also shows the fitted density-dependent component $\hat{\gamma}\, X_{t-1}$ in the form of a straight line. This plot suggests no gross departures from the fitted model, although there is a hint of a flattening of growth rates at the highest densities.

DISCUSSION

Long population time series are especially valuable in providing information about the sources of population variability. Unfortunately, such time series often have missing values, so that standard time series methods cannot be applied. It seems fair to say that, in this situation, the method of choice has been to fill in the missing values by some kind of interpolation scheme. As noted, this alters the statistical structure of the time series, although the consequences are unlikely to be severe when the proportion of missing values is low. However, when this is not the case, this method can lead to bias (Barnea 2004). The purpose of this paper has been to describe and illustrate how a population

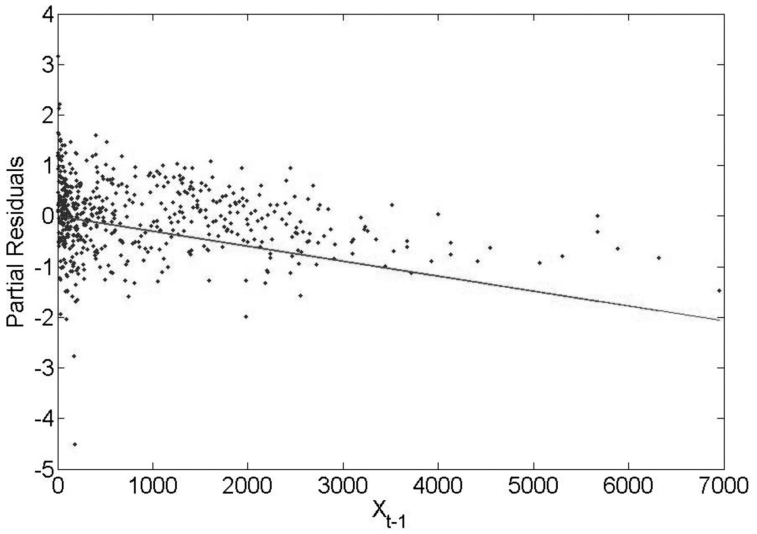

Fig. 4. Partial residuals ($R_t = \log X_t/X_{t-1}$) – ($\hat{\beta} + \hat{S}_t$) vs. X_{t-1}.

model can be fit by maximum likelihood in the presence of missing values (Matlab code for all computations available upon request).

The time series analysis shows that the *Peridinium* dynamics was extremely regular for at least 25 years and governed by a strong periodic signal. The erratic nature of the time series, which is certainly nontrivial (see Fig. 1), could be explained solely as noise superimposed on the underlying dynamics. This is an important point that need not have been the case, and indicates a very successful model fit. Furthermore, the analysis identified significant density dependence in the phytoplankton dynamics, which is interesting for a number of reasons. Firstly, the problem of detecting density dependence in ecological data sets is not an easy one. To our knowledge, the analysis here is the first of its kind for detecting density dependence in seasonally forced data sets. Secondly, from an ecological perspective it is interesting that density dependence was identified in this time series. Inspection of the time series by eye shows no indication of this process whatsoever. However, it seems reasonable that in the bloom period the phytoplankton are in fact competing for one or more limiting nutrients. Indeed, as simulations have shown, the model itself becomes unstable without the inclusion of a density-dependent term, attesting to the importance of this regulating process (Barnea, 2004). The method presented in this paper, together with other recent work (Clark and Bjørnstad, 2004), should be seen as a modest first step in developing new techniques for analyzing nonuniformly sampled time series.

ACKNOWLEDGMENTS

We thank Tamar Zohary and Utza Pollingher of the Yigal Alon Kinneret Limnological Laboratory for the *Peridinium* population data, Andy Beet for his help with programming, and Rony Braunstein for his critical comments. We are grateful for the support of the EU-Phytoplankton-On-Line grant.

REFERENCES

Barnea, O. 2004. Modelling nonuniformly sampled ecological time series. M.Sc. thesis, Tel Aviv University.

Berman, T., Pollingher, U. 1974. Annual and seasonal variations of phytoplankton, chlorophyll, and photosynthesis in Lake Kinneret. Limnol. Oceanogr. 19: 31–54.

Berman, T., Stone, L., Yacobi, Y.Z., Kaplan, B., Schlichter, M., Nishri, A., Pollingher, T. 1995. Primary production and phytoplankton in Lake Kinneret: a long-term record (1972–1993). Limnol. Oceanogr. 40: 1064–1076.

Clark, J.S., Bjørnstad, O.N. 2004. Population time series: process variability, observation errors, missing values, lags, and hidden states. Ecology 85: 3140–3150.

Den Boer, P.J., Reddingius, J. 1996. Regulation and stabilization paradigms in population ecology. Chapman & Hall, London, 397 pp.

Jassby, A.D., Powell, T.M. 1990. Detecting changes in ecological time series. Ecology 71: 2044–2050.

Kohn, R., Ansley, C.F. 1986. Estimation, prediction, and interpolation for ARIMA models with

missing data. J. Am. Statistical Assoc. 81: 751–761.
May, R.M. 1974. Stability and complexity in model ecosystems (2nd ed.). Princeton University Press, Princeton, NJ, 265 pp.
Pollingher, U., Serruya, C. 1976. Phased division of *Peridinium cinctum fa. Westii* and the development of blooms in Lake Kinneret. J. Phycol. 11: 155–162.
Solow, A.R., Stone, L., Rozdilsky, I. 2003. A critical smoothing test for multiple equilibria. Ecology 84: 1459–1463.
Steven, D.M., Glombitza, R. 1972. Oscillatory variation of a phytoplankton population in a tropical ocean. Nature 237: 105–107.