INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION

INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
34100 TRIESTE (ITALY) - P.O.B. 586 - MIRAMARE - STRADA COSTIERA 11 - TELEPHONES: 224281/2/3/4/5/6
CABLE: CENTRATOM - TELEX 460392-I

SMR/99- 11

AUTUMN COURSE ON MATHEMATICAL ECOLOGY

(16 November - 10 December 1982)

# ON THE ANALYSIS OF ECOLOGICAL TIME SERIES

L.J. GROSS

Mathematics Department
University of Tennessee
Knoxville, TN 37996
U.S.A.

ON THE ANALYSIS OF ECOLOGICAL TIME SERIES

Louis J. Gross

## INTRODUCTION:

As a science which is ultimately based on testing hypotheses, ecology makes use of a variety of statistical techniques to carry out analysis of data from field and laboratory observations. Aside from applications of standard statistical methods such as t-tests, F-tests, and analysis of variance, there have been very sophisticated applications of principle components analysis, factor analysis and other multivariate approaches, especially in the analysis of components structuring plant and animal communities (1). There exist whole volumes devoted to statistical ecology (2), and any overview at this area would require considerable time. Instead what I will attempt here is to give some insight into one area of statistical analysis that arises frequently in ecology, namely that of time series. For our purposes a time series is a collection of numerical observations arranged in a natural order, such as each observation being associated with a particular instant of time or point in space. The observations may consist of vector quantities in which case we have a multivariate time series. Some examples would include monthly total rainfall occurring at a particular weather station over several years, continuous measurements of any environmental variable (temperature, wind speed, irradiance, etc.) at any fixed location over a day, yearly population size estimates obtained from census data, chlorophyll counts of phytoplankton concentration along a transect in an ocean, and diversity measurements along a latitudinal



gradient.

Time series may be classified in several ways, one of which concerns whether the data is taken continuously in time (or space) or whether observations occur at discrete intervals. Continuous observations arise in communications and signal theory, but are somewhat rare in ecological practice. For a discrete time-series, the observations generally are taken over equally spaced intervals; the analysis of non-equally spaced data becomes considerably more complicated. A time series may be one- or multi-dimensional depending upon where the underlying ordering occurs. For example, the distribution of temperature throughout a lake could be described as a 4-dimensional time series. In practice, the analysis of multi-dimensional problems is very complicated (3), and I will here mainly deal with only the simpler, one-dimensional case. In any time series, successive observations are generally not independent, so one must take into account the time order of the data. When successive values are dependent, future values may be predicted from past values. If they can predict exactly, the series is deterministic. If the future is only partly determined by the past, the series is stochastic.

## OBJECTIVES OF TIME SERIES ANALYSIS:

The types of analysis performed on any particular time series will depend upon the objectives desired. These objectives may include (4):

(i) Description - We may desire to determine if there is a "trend" in the data and if so, determine its form. There may also be seasonal effects or other periodicities, and outliers - meaning data points which are far away from most.

(ii) Explanation - Here, if observations are taken on two or more variables, it may be possible to explain variations in one time series by those in another. This may involve setting up a model of the system and using the data to test the model.

(iii) Prediction - Given an observed time series, we may want to estimate future values. This may be done without requiring any attempt to explain the data as in (ii), by assuming the basic mechanism generating the data doesn't change through time, allowing statistical prediction methods to work. For example, we may want to estimate future population sizes, given data over several past years, without building a comprehensive model of the population dynamics, which could well require extensive biological knowledge of the species involved.

(iv) Control - Here we may wish the time series to have certain values in the future when we have control over part of the system. This is the usual problem faced by wildlife managers in setting hunting or fishing limits in order to maintain populations at the desired levels.

In actually analyzing a particular series, several points are worth considering. First it is necessary to decide what questions you wish to answer. Given this, the techniques which may be used to answer these questions should be determined. It is only after this that data should be taken, in order that it will indeed be done in an appropriate manner for your chosen techniques to apply. In practice, the data is often taken with little regard for the analysis techniques to be used, which can lead to great difficulty in actually performing the analysis. For example, many techniques work best with evenly-spaced data with no missing values, and it would be worth extra effort to obtain data in this form during the observations. Once the data is obtained, what is perhaps the simplest, most useful technique is to merely graph it. Often this will serve to guide further analysis.



TECHNIQUES OF ANALYSIS:

One way of structuring the field of time series analysis is by considering these techniques which have been most useful. As with any subject of analysis, however, it is preferable to keep an open mind about which techniques are most appropriate for your purposes, rather than sticking to one approach no matter what. Time series analysis is indeed a field in which breadth of knowledge of many techniques is essential. Several excellent texts are available to cover these topics (4,5,6), and I here give only a brief overview of the techniques which have been the most useful in ecological applications (7).

TIME DOMAIN ANALYSIS:

The first techniques I discuss concern analysis of the data as ordered by time (or space), in contrast to frequency domain analysis which utilizes a transformation to frequency-space. One of the prime objectives of time domain analysis is to determine trends in the data, which we define as a long term change in the mean. One must be careful about trend analysis since what is a trend in a short time series may turn out to be part of an oscillation when a longer time series is available. To measure trends it is often appropriate to do a curve fit, or regression, so that if $x_t$, $t = 0,1,\ldots,N$ are the data points, and $y_t = f(t)$ is the regression curve, then the parameters in $f(t)$ are chosen to minimize $E = \sum_{t=0}^{N} (y_t - x_t)^2 w(t)$ where $w(t)$ is a weighting function. The simplest case is a linear regression in which $y_t = at + b$. Far more complicated forms may be used however, see for example (8) in which a sum of an exponential and four sinusoids is fit to data on atmospheric $CO_2$

content. Often the choice of $y_t$ is an iterative process with successively better regression curves chosen until the least-squares error E is below some acceptable level.

A variety of techniques to either display a trend or remove it, involve the use of filters. The simplest of these is the linear filter, which is a new time series given by

$$y_t = \sum_{r=-q}^{s} a_r x_{t+r} .$$

Here the $a_i$'s are chosen to meet some desired goal. In the case $\sum a_r = 1$, this is called a moving average, which may be thought of as smoothing out local fluctuations and estimating the local mean. Another type of filter is obtained by differencing, e.g. $y_t = x_{t+1} - x_t = \nabla x_t$. This differencing may be repeated a number of times if desired, and is typically used to remove trends. The above smoothing operations cause data to be lost at the beginning and end of the series. One alternative is to use an asymmetric filter

$$y_t = \sum_{j=o}^{q} a_j x_{t-j} .$$

The case of exponential smoothing takes $q = \infty$ and $a_j = \alpha(1-\alpha)^j$. It is important to keep in mind that any filter will cause certain frequencies in the data to be emphasized and others to be deleted.

Some definitions now are necessary:

Defn. The autocovariance coefficient at lag k of a time series $x_t$ is

$$c_k = \frac{1}{N} \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$$



where $\bar{x} = \frac{1}{N} \sum_{t=1}^{N} x_t$ is the mean of the series. Note that $c_0$ is the sample variance of the series.

Defn. The autocorrelation coefficient at lag k is

$$r_k = \frac{c_k}{c_0} .$$

Here $r_k$ measures the degree to which observations a distance k apart are correlated. It is the same as the usual correlation coefficient, only calculated using the same string of data. The correlogram is a plot of $r_k$ vs. k and the analysis of correlograms is often the heart of analysis in the time domain. The approach is to first estimate the trend and seasonal effects, subtract them out so that all that is left are residuals, then analyze the residuals to see if they fit any particular probability model. I now discuss some of the standard models.

Let $\{X_t\}$, $t = 0,1,2,\ldots$ be a stochastic process on a probability space $(\Omega, \mathcal{F}, P)$. We define the mean function, $\mu(t) = E[X_t]$, the variance function $\sigma^2(t) = E[(X_t - E[X_t])^2]$, and the autocovariance function

$$\gamma(t_1,t_2) = E[(X_{t_1} - \mu(t_1))(X_{t_2} - \mu(t_2))] .$$

Defn. A stochastic process $\{X_t\}$ is second order (or covariance or weakly) stationary if $E[X_t] = \mu$, $E[X_t^2] < \infty$ and $\gamma(t, t+\tau) = \gamma(\tau)$, i.e. the mean and variance are time-independent and the covariance depends only upon the lag. The autocorrelation function for a weakly stationary process is

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} .$$

For our purposes, the most important property of stationary processes is that they are ergodic:

Thm. (Weak Ergodic) (9) Let $\{X_t\}$ be weakly stationary and let $\overline{X}_t = \frac{1}{t}(X_1 + X_2 + \ldots + X_t)$ . Then there exists a random variable $\overline{X}$ such that

$$\lim_{t \to \infty} E[(\overline{X}_t - \overline{X})^2] = 0 .$$

In addition $\overline{X}$ is constant, $\overline{X} = E[X_1]$ , if and only if

$$\lim_{N \to \infty} \frac{1}{N} \sum_{v=0}^{N-1} \gamma(v) = 0 .$$

Under the conditions of this theorem, we can estimate the mean of the process $\{X_t\}$ by using time averages, rather than having to take several different realizations (or time series) of the process. Similarly, if the process is Gaussian, we can estimate the covariance function $\gamma(\tau)$ from a single time series. Ergodicity refers to the equivalence of time averages and ensemble or sample averages. A similar theorem giving convergence almost surely rather than in mean square requires strict stationarity of the process.

## SOME PROBABILITY MODELS:

Some of the typical models include:

1. Purely random process. Let $Z_t$ , $t = 1,2,\ldots$ be a sequence of mutually independent, identically distributed random variables. Then $E[Z_t] = \mu \ \forall t$ and $\gamma(k) = \begin{cases} 0 & \text{for } k > 0 \\ \sigma^2 & \text{for } k = 0 \end{cases}$ where $\sigma^2 = Var(Z_k)$ .

Thus $\{Z_k\}$ is weakly stationary - it is sometimes called discrete white noise.



2. Moving average process. Let $\{Z_t\}$ be a purely random process with mean 0 and variance $\sigma^2$ . Then $X_t$ is a moving average process of order $m$ if

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \ldots + \beta_m Z_{t-m}$$

where $\{\beta_i\}$ are constants. Then $E[X_t] = 0$ , $Var(X_t) = \sigma^2 \sum_{i=0}^{m} \beta_i^2$ and

$$\gamma(k) = \begin{cases} 0 & k > m \\ \sigma^2 \sum_{i=0}^{m-k} \beta_i \beta_{i+k} & k = 0,1,2,\ldots,m \\ \gamma(-k) & k < 0 . \end{cases}$$

Thus $\{X_t\}$ is weakly stationary.

3. Autoregressive process. Let $\{Z_t\}$ be purely random, mean 0 , variance $\sigma^2$ . Then $X_t$ is autoregressive of order $m$ if

$$(*) \qquad X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_m X_{t-m} + Z_t .$$

Here $X_t$ is regressed on past values of the process. It is clear that $E(X_t) = 0$ , and to derive conditions for stationarity, assume $X_t$ is stationary, multiply (*) by $X_{t-k}$ , $k > 0$ , take expectations, divide by $Var(X_t)$ to get

$$\rho(k) = x_1 \rho(k-1) + \ldots + \alpha_m \rho(k-m) \quad \text{for all } k > 0 .$$

(Yule-Walker equations). It can be shown that $\{X_t\}$ is stationary if $\rho(k) \to 0$ as $k \to \infty$ , which will hold if all roots of $y^m - \alpha_1 y^{m-1} - \alpha_2 y^{m-2} - \ldots - \alpha_m = 0$ are less than 1 in magnitude. Bulmer (10) uses a first order autoregressive model with a sinusoidal trend term to analyze cycles of snowshoe hare and lynx.

4. Mixed autoregressive-moving average model (ARMA). Here $\{Z_t\}$ is purely random, and

$$X_t = \alpha_1 X_{t-1} + \dots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \dots + \beta_q Z_{t-q}$$

where $\{\alpha_i\}$ and $\{\beta_i\}$ are constants. This type of model has been used to estimate parameters in a linear compartment ecosystem model (11). See (6) for its use in forecasting.

5. Autoregressive, integrated moving average model (ARIMA). We take the $d^{th}$-differences

$$W_t = \nabla^d X_t = \nabla (\nabla \dots \nabla X_t) ,$$

then the model is

$$W_t = \alpha_1 W_{t-1} + \dots + \alpha_p W_{t-p} + Z_t + \beta_1 Z_{t-1} + \dots + \beta_p Z_{t-q} .$$

The use of differences here causes the original series to be detrended. ARIMA is extremely useful in prediction problems (6). However, there can be errors developed since this is a linear model and the differencing may not adequately remove the non-stationarity of a non-linear model (12). Successful results on forecasting mosquito populations have been obtained (13).

A general approach to estimation in the time domain (here estimation refers to deciding a proper model and its parameters) is to detrend the series, construct the correlogram, and then compare the correlogram to those from a variety of models of the above type, to see which best fits the data. This may well be an iterative process, and a variety of approaches to estimating parameters once a model is chosen are available (6). Since all the above models are linear, they



have limitations, though some extensions of them to non-linear cases have been made (14).

## ANALYSIS IN THE FREQUENCY DOMAIN:

As a counterpart to the above time domain analysis, we may consider the various frequency components which make up a particular time series. This is called spectral analysis, and is based upon Fourier series techniques. If a time series appears to have periodicities at certain known frequencies, then it might be modeled as

$$X_t = \sum_{i=1}^{n} R_i \cos(w_i t + \theta_i) + Z_t$$

where $\{Z_t\}$ are i.i.d. random variables. See (8) for an application to striped bass catches. The general analysis follows from:

Theorem (9). For a real-valued weakly stationary stochastic process with autocovariance $\gamma(k)$ , there is a monotone increasing function $F(w)$ , called the spectral distribution function, such that

$$\gamma(k) = \int_{-\pi}^{\pi} \cos wk \; d(F(w)) , \quad F(w) = 1 - F(-w) .$$

Here $\gamma(0) = \sigma^2 = \int_0^{\pi} dF(w)$ , so we may think of $F(w)$ as the contribution to the variance of the process caused by frequencies in the range $(0,w)$ . If $F(w)$ is differentiable, its derivative $f(w)$ is the spectral density function, or power spectrum, and

$$f(w) = \frac{1}{\pi} \sum_{k=-\infty}^{\infty} \gamma(k)\, e^{-iwk} .$$

Thus, for a stationary process, the power spectrum may be estimated once the autocovariance is known.

Because a time series consists of discrete sample points, it's obvious one cannot expect to see variations in the data which occur on a scale smaller than the sampling interval. If the interval between samples is $\Delta t$ , then the highest frequency we can see is $f_N = \frac{1}{2\Delta t}$ , the Nyguist frequency. Similarly, since we only sample over a finite time, we cannot see cycles which are on a time scale longer than the total sampling interval. These facts may put severe limitations on what time scales we can detect and so should be taken into account before the data is collected. In general, the determination of sampling intervals for ecological systems is quite difficult (15).

In order to obtain a consistent estimate of the power spectrum, one technique is to transform the truncated autocovariance (5), to obtain

$$\hat{f}(w) = \frac{1}{\pi} \{\lambda_0 c_0 + 2 \sum_{k=1}^{M} \lambda_k c_k \cos wk .$$

Here $\{\lambda_k\}$ are a set of weights called the lag window and $M < N$ is the truncation point. A variety of choices of the lag window exist, along with other estimators for $f(w)$ . Spectral analysis techniques have been extensively applied in phytoplankton studies (16) and are often used in an iterative approach with time-domain approaches to detrend data (8).



REFERENCES


1. Gauch, Hugh G., Jr. 1982. Multivariate Analysis in Community Ecology. Cambridge Univ. Press, N.Y.

2. Patil, G.P. (editor). 1971 on. Statistical Ecology Series. Volume 1-13. International Cooperative Publishing House, Burtonsville, Maryland.

3. Bennett, R.J. 1979. Spatial Time Series: Analysis - Forecasting - Control. Pion Limited, London.

4. Chatfield, C. 1975. The Analysis of Time Series: Theory and Practice. Halsted Press.

5. Bloomfield, P. 1976. Fourier Analysis of Time Series: An Introduction. Wiley, NY.

6. Box, G. and G. Jenkins. 1970. Time Series Analysis - Forecasting and Control. Holden-Day, N.Y.

7. Shugart, H.H., Jr. (editor). 1978. Time Series and Ecological Processes. SIAM. Philadelphia, PA.

8. Rust, B.W. and B.L. Kirk. 1978. Inductive Modelling of Population Time Series. P. 156-192 in H.H. Shugart, Jr. (editor) ibid.

9. Karlin, S. and H.M. Taylor III. 1975. A First Course in Stochastic Processes. Academic Press, NY.

10. Bulmer, M.G. 1975. The Statistical Analysis of the Ten Year Cycle. P. 141-153 in H.H. Shugart (editor) ibid.

11. Stevens, D.L. and W.S. Overton. 1975. An approach to estimating the parameters of a forced linear system using an ARMA model. p. 19-106 in H.H. Shugart (editor) ibid.

12. van der Vaart, H.R. 1975. An example of the performance of time series methods with respect to a known model. p. 107-121 in H.H. Shugart (editor) ibid.

13. Hacker, C.S. 1975. Autoregressive and transfer function models of mosquito populations. p. 294-303 in H.H. Shugart (editor) ibid.

14. Granger, C.W.J. 1975. Some new time series models: non-linear, bilinear and non-stationary. p. 122-138 in H.H. Shugart (editor) ibid.

15. Mulholland, R.J. and C.M. Gowdy. 1975. Determination of sampling intervals for ecological processes. P. 71-86 in H.H. Shugart (editor) ibid.

16. Platt, J. and K. Denmann. 1975. Spectral Analysis in Ecology Ann. Rev. Ecol. System. 6: 189-210.