Chapter

12

Ecological data series

12.0 Ecological series

The use and analysis of *data series* has become increasingly popular in ecology, especially because many terrestrial, aquatic and atmospheric observing stations measure and record environmental variables either automatically or with human intervention. Ecological data series contain continuous or discrete (discontinuous) variables sampled over time or along transects in space.

Stochastic process

A data series is a sequence of observations that are *ordered* along a temporal or spatial axis. As mentioned in Section 1.0, a series is one of the possible realizations of a *stochastic process*. A *process* is a phenomenon (response variable), or a set of phenomena, which is organized along some independent axis. In most cases, the independent axis is time, but it may also be space, or a trajectory through both time and space (e.g. sampling during a cruise). *Stochastic processes* generally exhibit three types of components, i.e. deterministic, systematic, and random. Methods for the numerical analysis of data series are designed to characterize the deterministic and systematic components present in series, given the probabilistic environment resulting from the presence of random components.

The most natural axis along which processes may be studied is *time* because temporal phenomena develop in an irreversible way, and independently of any decision made by the observer. The temporal evolution of populations or communities, for example, provides information that can unambiguously be interpreted by ecologists. Ecological variability is not a characteristic limited to the time domain, however; it may also be studied across space. In that case, the decisions to be made concerning the observation axis and its direction depend on the working hypothesis. In ecology, the distinction between space and time is not always straightforward. At a fixed sampling location, for example, time series analysis may be used to study the spatial organization of a moving system (e.g. migrating populations, plankton in a current), whereas a spatial series is required to assess temporal changes in that same

Eulerian
Lagrangian

system. The first approach (i.e. at a fixed point in space) is called *Eulerian*, and the second (i.e. at a fixed point within a moving system) is known as *Lagrangian*.

Periodic phenomena

Ecologists are often interested in *periodic* changes. This follows in part from the fact that many ecological phenomena are largely determined by geophysical rhythms, but there are also rhythms that are endogenous to organisms or ecosystems. The geophysical cycles of glaciations, for example, or, at shorter time scales, the solar (i.e. seasons, days) or lunar (tides) periods, play major roles in ecosystems. Endogenous rhythms, also called biological clocks (including the well-known circadian, i.e. 24-hour, rhythms), are extensively described in the scientific literature.

The analysis of data series often provides unique information concerning ecological phenomena. However, the quality of the results depends to a large extent on the *sampling design*. As a consequence, data series must be sampled following well-defined rules, in order (1) to preserve the spatio-temporal variability, which is often minimized on purpose in other types of ecological sampling design, and (2) to take into account the various conditions prescribed by the methods of numerical analysis. These conditions will be detailed later in the present chapter. An even more demanding framework prevails for *multidimensional series*, which result from sampling several variables simultaneously. Most numerical methods require that the series be made up of *large numbers of observations* ($n > 100$, or even $n > 1000$) for the analysis to have enough statistical power to provide conclusive results, especially when large random variation is present. Long series require extensive sampling. This is often carried out, nowadays, using equipment that automatically measures and records the variables of ecological interest. There are also a few methods that have been especially designed for the analysis of short time series; they are discussed below.

Observational window
Lag
Period
Frequency
Wavelength
Wavenumber

The most fundamental constraint in periodic analysis is the *observational window*. The width of this window is determined by the number of observations in the data series (n) and the interval (time or distance) between successive observations. This interval is called the *lag*, Δ; for the time being, it is assumed to be uniform over the whole data series. These two characteristics set the time or space domain that can be "observed" when analysing data series (Table 12.1). For temporal data, one refers to either the *period* (T, in time units) or the *frequency* ($f = 1/T$) whereas, for spatial data, the corresponding concepts are the *wavelength* (λ, in spatial distance units) and the *wavenumber* ($1/\lambda$).

The length of the series (Δn) sets, for temporal data, the *fundamental period* ($T_0 = \Delta n$) or *fundamental frequency* ($f_0 = 1/T_0 = 1/\Delta n$) and, for spatial data, the *fundamental wavelength* ($\lambda_0 = \Delta n$) or *fundamental wavenumber* ($1/\lambda_0 = 1/\Delta n$). *Harmonic periods* and *wavelengths* are *integral fractions* of the fundamental period and wavelength, respectively ($T_i = T_0/i$ and $\lambda_i = \lambda_0/i$, where $i = 1, 2, \ldots n$), whereas *harmonic frequencies* and *wavenumbers* are *integral multiples* of the fundamental frequency and wave number, respectively ($f_i = if_0$ and $1/\lambda_i = i/\lambda_0$). Concerning the actual limits of the observational window, the *longest* period or wavelength that can be statistically investigated is, at best, equal to *half the length* of the series ($\Delta n/2$). For

Table 12.1 Characteristics of the observational window in periodic analysis. Strictly speaking, the length of a data series is $(n-1)\Delta$ but, for simplicity, one assumes that the series is long, hence $(n-1) \approx n$.

Harmonic i	Period (T_i) Wavelength (λ_i)	Frequency (f_i) Wavenumber (i)	
1	$n\Delta$	$1/n\Delta$	Fundamental value, i.e. the whole series
2	$n\Delta/2$	$2/n\Delta$	Limit of observational window
.	.	.	
.	.	.	
i	$n\Delta/i$	$i/n\Delta$	ith harmonic
.	.	.	
.	.	.	
$n/2$	2Δ	$1/2\Delta$	Limit of window: Nyquist frequency

example, in a study on circadian (24-h) rhythms, the series must have a *minimum* length of two days (better 4 days or more). Similarly, in an area where spatial structures are of the order of 2 km, a transect must cover *at least* 4 km (better 8 km or more). Similarly, the *shortest* period or wavelength that can be resolved is equal to *twice the interval* between observations (2Δ). In terms of frequencies, the highest possible frequency that can be resolved, $1/2\Delta$, is called the *Nyquist frequency*. For example, if one is interested in hourly variations, observations must be made *at least* every 30 min. In space, in order to resolve changes at the metre scale, observations must be collected along a transect *at least* every 50 cm, or closer.

Nyquist frequency

To summarize the above notions concerning the observational window, let us consider a variable observed every month during one full year. The data series would allow one to study periods ranging between (2×1 month = 2 months) and (12 months/2 = 6 months). Periods shorter than 2 months and longer than 6 months are outside the observational window. In other words, statistical analysis cannot resolve frequencies higher than 1/(2 months) = 0.5 cycle month^{-1} = 6 cycles year^{-1} (Nyquist frequency), or lower than 1/(6 months) = 0.167 cycle month^{-1} = 2 cycles year^{-1}. The longest period (or lowest frequency) of the observational window is easy to understand, by reference to the usual notion of degrees of freedom (Box 1.2). Indeed, in order to have minimum certainty that the observed periodic phenomenon is real, this phenomenon must be observed at least twice, which provides only one degree of freedom. For example, if an annual cycle was observed over a period of one year

only, there would be no indication that it would occur again during a second year (i.e. no degree of freedom). A similar reasoning applies to the shortest period (or highest, Nyquist frequency) detectable in the observational window. For example, if the observed phenomenon exhibits monthly variation (e.g. oscillations between maximum and minimum values over one month), two observations a month would be the absolute minimum required for identifying the presence of that cycle.

Most methods described in the present chapter are limited to the observational window. However, some methods are mathematically capable of going beyond the upper limit (in terms of periods) of the window, because they can fit incomplete cycles of sine and cosine functions to the data series. This is the case of Dutilleul's modified periodogram (Section 12.4) and spectral analysis (Section 12.5). A significant period found in this region (e.g. a 3-month period in a data series 4 months long) should be interpreted with care. It only indicates that a longer time series should be observed and analysed (e.g. > 1 year of data) before drawing ecological conclusions.

Aliasing

There exists another constraint, which is also related to the observational window. This constraint follows from a phenomenon known as *aliasing*. It may happen that the observed variable exhibits fluctuations whose frequency is *higher than the Nyquist frequency*. This occurs when a period T or wavelength λ of the observed variable is smaller than 2Δ. Undersampling an important high-frequency fluctuations may generate an artificial signal in the series, whose frequency is *lower than the Nyquist frequency* (Fig. 12.1). Researchers unaware of the phenomenon could attempt to interpret this artificial low frequency in the series; this would obviously be incorrect. To avoid aliasing, the sampling design must provide at least four data points per cycle of the *shortest* important period or wavelength of the variable under study. The latter period or wavelength may be determined either from theory or from a pilot study.

The sections that follow explore various aspects of series analysis. The methods discussed are those best adapted to ecological data. Additional details may be found in the biologically-oriented textbook of Diggle (1990) and the review paper of Fry *et al*. (1981), or in other textbooks on time series analysis, e.g. Jenkins & Watts (1968), Bloomfield (1976), Box & Jenkins (1976), Brillinger (1981), Priestley (1981a, b), Kendall *et al*. (1983), Chatfield (1989), Kendall & Ord (1990), Venables & Ripley (2002), Dutilleul (2011) and Shumway & Stoffer (2011, with R examples). Methods for analysing time series of ecological and physiological chronobiological data were reviewed by Legendre & Dutilleul (1992).

12.1 Characteristics of data series and research objectives

Signal
Trend
Noise

Observed data series may be decomposed into various components, which can be studied separately since they have different statistical and ecological meanings. Figure 12.2 shows an artificial data series constructed by adding three components: a periodic signal, a trend, and a noise component. Series may be analysed in terms of

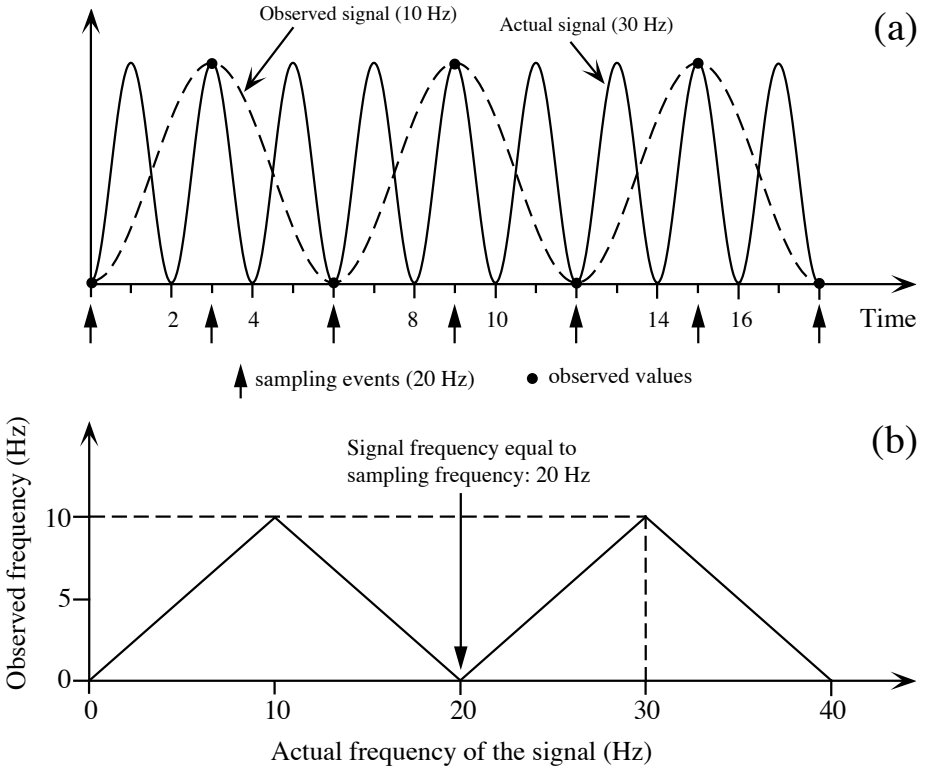


Figure 12.1 Aliasing. (a) The artificial signal detected in the data series (dashed line) is caused by observations (dots) made at a frequency lower than twice that present in the series under study (solid line). Along the abscissa, 1 time unit = 1/60 s. (b) With a sampling frequency of 20 Hz, the observed frequency (ordinate) varies between 0 and 10 Hz, as the actual frequency of the signal increases (abscissa). The observed frequency would be equal to the frequency of the signal (no aliasing) only for a signal ≤ 10 Hz, which is half the 20 Hz sampling frequency. In the example, the frequency of the signal is 30 Hz and the observed (aliased) frequency is 10 Hz (dashed line).

deterministic change (trend), *systematic* (periodic) variability, and *random* fluctuations (noise). Data series may be recorded with different objectives in mind (Table 12.2), to which are associated different methods of time series analysis. The following presentation of objectives is largely drawn from Legendre & Dutilleul (1992).

Trend

Objective 1. — Ecological data series often exhibit a deterministic component, known as the *trend*. The trend may be linear, polynomial, cyclic, etc. This deterministic component underlies the evolution of the series (Fig. 12.2a). It must be extracted as the first step of the analysis.

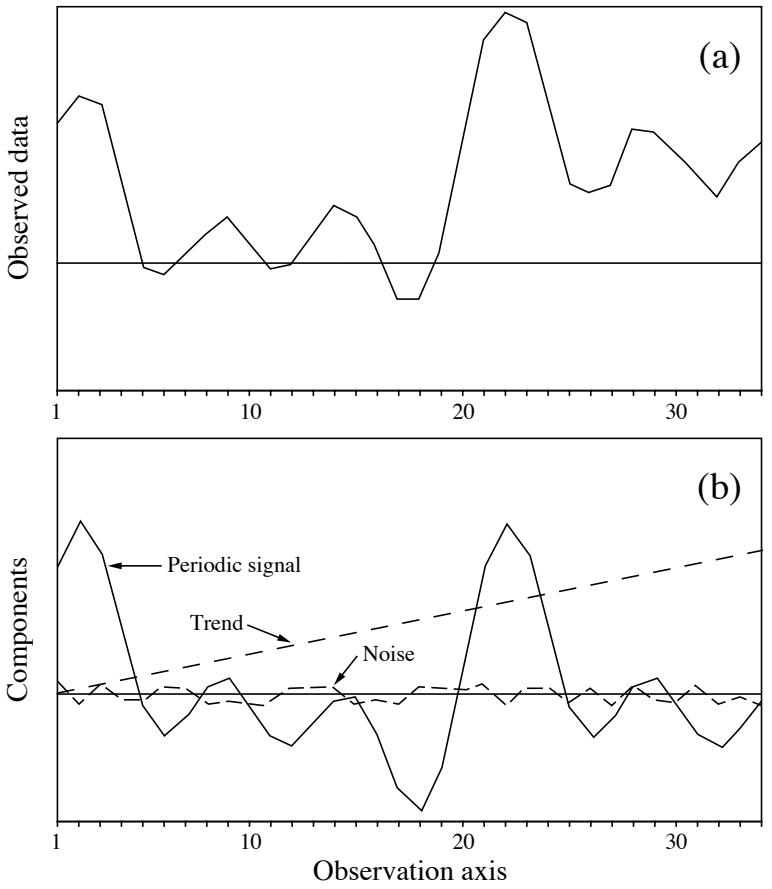


Figure 12.2 Artificial data series (a) constructed by adding the three components shown in (b), i.e. a periodic signal and a noise component, whose combination gives a stationary series (not illustrated), and a linear trend. The periodic signal is the same as in Fig. 12.13. There are $n = 34$ data points sampled at regular intervals. The overall mean of the noise signal is zero, by definition.

In some cases, determining the *trend* is the chief objective of the study. For example, progressive changes in the characteristics of an ecosystem, over several years, may be used to assess whether this system is responding or not to anthropogenic effects. In such a case, the problem would be to characterize the long-term trend, so that the annual cycle as well as the high-frequency noise component would be of no interest. Long-term trends in data series may be modelled by regression (Section 10.3). Linear regression is used when the trend is (or seems to be) linear. In other cases, the ecological hypothesis or a preliminary examination of the series may indicate that the trend is of some other mathematical form (e.g. logistic), in which case the methods of polynomial or nonlinear regression should be used (e.g. Ross, 1990).

Stationarity

Detrending
Trend extraction

In contrast, ecologists primarily interested in the *periodic* component of data series (Objective 2) consider the long-term trend as a nuisance. Even when the trend is not of ecological interest, it must be extracted from the series because most methods of analysis require that the series be *stationary*, i.e. that the mean, variance, and other statistical properties of the distribution be constant over the series. In the numerical example of Fig.12.2, the observed data series (a) is obviously not stationary. It becomes so if the linear trend shown in (b) is removed by subtraction; this operation is called *detrending* (or *trend extraction*). The trend may be estimated, in this case, by linear regression over a reasonably long segment of data; detrending consists in calculating the regression residuals. In practice, the analysis of series only requires *weak*, or *second-order*, or *covariance stationarity*, i.e. the mean and variance are constant along the series and the autocovariance (or autocorrelation) function depends only on the distance between observations along the series; two observations separated by a given interval have the same autocovariance no matter where they occur in the series. Extracting trends may be done in various ways, which are detailed in Section 12.2.

Some *low-frequency periodic* components may also be considered as trends, especially when these are both trivial and known *a priori* (e.g. an annual cycle). A long-term trend as well as broad-scale periodic components may be extracted in order to focus the analysis on finer components of the data series. Again, regression or other statistical methods (Section 12.2) may be used to model the low-frequency components and compute residuals on which the analysis could be carried out.

Objective 2. — Identifying *characteristic periods* is a major objective of series analysis in ecology. It is generally done for one variable at a time, but it is also possible to study multidimensional series (i.e. several variables, often analysed two at a time). Ecological series always exhibit irregular and unpredictable fluctuations, called *noise* (Fig. 12.2b), which are due to non-permanent perturbation factors. The larger the noise, the more difficult it is to identify characteristic periods when analysing stationary series. Table 12.3 summarizes the methods available to do so; several of these are described in Sections 12.3 to 12.5.

Objective 3. — One method for identifying characteristic periods is spectral analysis. In this analysis, the variance of the data series is partitioned among frequencies (or wavenumbers) in order to estimate a *variance spectrum*. Section 12.5 shows that the spectrum is a global characteristic of the series, and presents examples where the spectra are interpreted as reflecting ecological processes.

Objective 4. — There are data series that do not behave in a periodic manner. This may be because only one or even part of a cycle has been sampled or, alternatively, because the variables under study are not under the control of periodic processes. Such series may exhibit structures other than periodic, along time or a spatial direction. In particular, one may wish to identify *discontinuities* along multidimensional data series. Such discontinuities may, for example, characterize *ecological succession*. A commonly-used method for finding discontinuities is cluster analysis. To make sure that the multidimensional series gets divided into blocks, each one containing a set of

Table 12.2 Analysis of data series: research objectives and related numerical methods. Adapted from Legendre & Legendre (1984b) and Legendre & Dutilleul (1992).

Research objective	Numerical methods
1) Characterize the trend	• Regression (linear or polynomial)* • Moving averages • Variate difference method
2) Identify characteristic periods	→ Details in Table 12.3
3) Characterize series by spectrum	• Spectral analysis
4) Detect discontinuities in multivariate series	• Clustering the data series (with or without constraint) • Hawkins & Merriam or Webster segmentation methods
5) Correlate variations in a series with changes in other series	
5.1) Univariate target series	• Regression*: simple / multiple linear, nonlinear, splines • Cross-correlation
5.2) Multivariate target series	• Canonical analysis** • Mantel test*
6) Formulate a forecasting model	• Box-Jenkins modelling

Methods described in * Chapter 10 or ** Chapter 11.

temporally contiguous observations, authors have advocated to constrain clustering algorithms so that they are forced to only group observations that are contiguous. Various methods to do so are discussed in Section 12.6.

Objective 5. — Another objective is to *correlate variations* in the data series of interest (i.e. the *target* or *response variable*) with variations in series of some potentially *explanatory variable(s)*, with a more or less clearly specified model in mind. There are several variants. (1) When the sampling interval between observations is large, the effect of the explanatory variables on the target variable may be considered as instantaneous. In such a case, various forms of regression analysis may be used. When no explicit model is known by hypothesis, spline regression may be used to describe temporal changes in the target variable as a function of another variable (e.g. Press *et al*., 2007). These methods are described in Section 10.3. (2) When the interval between consecutive data is short compared to the periods in the target variable, it is sometimes assumed that the target variable responded to events that occurred at some previous time, although the exact delay (*lag*) may not be known. In such a case, the method of cross-correlation may be used to identify the time lag that maximises the correlation between the explanatory and target variables (Section 12.3). When the optimal lag has been found for each of the explanatory variables in a model, multiple regression can then be used, each explanatory variable being lagged by the

Table 12.3 Analysis of data series: methods for identifying characteristic periods. The approaches best suited to *short* data series are: the contingency periodogram, Dutilleul's modified periodogram, and maximum entropy spectral analysis. Adapted from Legendre & Legendre (1984b) and Legendre & Dutilleul (1992).

Type of series	Methods	
	Quantitative variables only	All precision levels
1) A single variable	• Autocorrelogram	• Spatial correlogram* (quantitative, qualitative)
	• Periodograms (Whittaker & Robinson, Schuster, Dutilleul)	• Contingency periodogram for qualitative data
	• Spectral analysis	• Kedem's spectral analysis for binary data
2) Two variables	• Parametric cross-correlation	• Nonparametric cross-correlation
	• Coherence and phase spectra	• Cross-contingency analysis
3) Multivariate series	• Multivariate spectral analysis	• Multivariate variogram*, Mantel correlogram*

* Methods described in Chapter 13.

appropriate number of sampling intervals. (3) The previous cases apply to situations where there is a single target variable in the series under study. When there are several target variables, the target series is multivariate; the appropriate methods of data analysis are globally called canonical analysis (Chapter 11). Two forms are of special interest here: redundancy analysis and canonical correspondence analysis. (4) Finally, the relationship between two distance matrices based on two multivariate data sets can be analysed using the Mantel test or its derived forms (Section 10.5 and Subsection 13.1.6) when the question strictly concerns distances.

Objective 6. — A last objective is to formulate a model to *forecast* the future behaviour of the target series. Following the tradition in economics, one way of doing that is to model the data series according to its own past behaviour (Section 12.7).

Testing for the presence of trends

The first problem encountered when analysing data series is to decide whether a *trend* is present or not. Visual examination of the series, which may be combined with previous knowledge about the process at work, is often sufficient to detect one or several trends. These may be monotonic (e.g. gradient in latitude, altitude, or water depth) or not (e.g. daily, lunar, or annual cycles). Four methods can be used to test for the presence of trends (extraction of trends: see Section 12.2).

• 1. The most widely used method is to regress the response data series **y** on the time variable. A significant regression coefficient indicates the presence of a linear trend, either positive or negative, in the series. Researchers must beware of a situation where

a trend is sought in a series that contains high variability nested into the series. For example, when looking for a trend among years in a series **y** covering several years, linear regression of **y** on the variable *years* may fail to detect a significant trend if the variation among months is high. To circumvent that problem, one can use a qualitative variable (or *factor*) coding for the months as covariable in the analysis. In practice, one can simply compute a linear model of **y** as a function of the quantitative variable *years* and the factor *months*, and check if the regression coefficient associated with *years* is significant.

• 2. The numbers of positive and negative *differences between successive values* in the series are counted. These are then subjected to a *sign test* (Table 5.2), where the null hypothesis (H_0) is that the plus and minus signs correspond to a population in which the two signs are present in equal proportions. Rejecting H_0 is indication of a trend.

• 3. All values in the series are *ranked* in increasing (or decreasing) order. *Kendall's rank correlation coefficient* (τ) (Subsection 5.3.2) may be used to assess the degree of resemblance between the rank-ordered series and the original one; this is done by computing the Kendall correlation between the original data series and the observation rank labels: $1, 2, 3, \ldots, n$. When τ is significantly different from zero, one can conclude that the series exhibits a *monotonic* trend. These two methods are described in Kendall & Ord (1990, pp. 21-22). The approach based on Kendall's τ is preferable to the sign test because it uses the actual data in the series instead of the differences between neighbouring values.

Up and down runs test

• 4. A nonparametric test, called the *up and down runs* test, is well suited to detect the presence of various types of trends. Consider again n values and, for each one, the sign of the difference from the previous value. The $(n - 1)$ signs would all be the same if the observations were monotonically increasing or decreasing. Cyclical data, on the other hand, would produce more long *runs* of "+" or "–" signs than expected for random data, or more short *runs*, depending on the sampling frequency within each cycle. A *run* is a set of like signs, preceded and followed (except at the end of the series) by opposite signs. Count the *number of runs* in the data series, including those of length 1 (e.g. a single "+" sign, preceded and followed by a "–"). The up and down runs test, described for instance in Sokal & Rohlf (1995), compares this number to the number of runs expected from a same-length sequence of random numbers.

When there is a *trend* in the series, it must be extracted using one of the methods discussed in Section 12.2. If, after detrending, the mean of the series is still not stationary, a second trend must be searched for and removed. When the series does not exhibit any trend, or after detrending, one must decide, before looking for periodic variability (Sections 12.3 to 12.5), whether the stationary series presents some kind of systematic variability or if, on the contrary, it simply displays the kind of variation expected from a random process. In other words, one must test whether the series is simply *random*, or if it exhibits *periodic variability* that could be analysed.

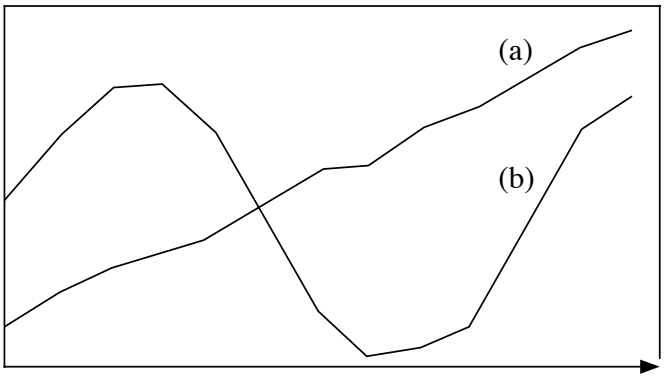


Figure 12.3 Two artificial series; (a) would be random if the linear trend was extracted, whereas (b) displays a cyclic trend.

Test series randomness

In some instances, as in Fig. 12.3, it is useless to conduct sophisticated tests, because the random or systematic character of the series is obvious. Randomness of a series may be tested as follows: identify the *turning points* (i.e. the peaks and troughs) in the series and record the distribution of the *number of intervals* (*phase length*) between successive turning points. It is possible to test whether these values correspond or not to those of a random series (Kendall & Ord, 1990, p. 20). This procedure actually tests the same null hypothesis as the up and down runs test described above. In practice, any ecological series with an average phase longer than two intervals may be considered non-random.

The overall procedure for analysing data series is summarized in Fig. 12.4. The following sections describe the most usual methods for extracting trends, as well as various approaches for analysing stationary series. It must be realized that, in some instances, variations in stationary series may be so small that they cannot be analysed, because they are of the same order of magnitude as the background noise.

If parametric statistical tests are to be conducted during the course of the analysis, *normality* must be checked (Section 4.6) and, if the data are not normally distributed, they must be *transformed* as explained in Subsection 1.5.6. In addition, several of the methods discussed in the following sections require that observations in the series be *equally spaced*. If they are not, data may be eliminated to make them *equispaced*, or else, missing data may be estimated by regression or other interpolation methods (Section 1.6); most methods of series analysis cannot handle missing values. Obviously, it is preferable to consider the requirement of equispaced data when designing a sampling program than to have to modify the data at the stage of analysis.

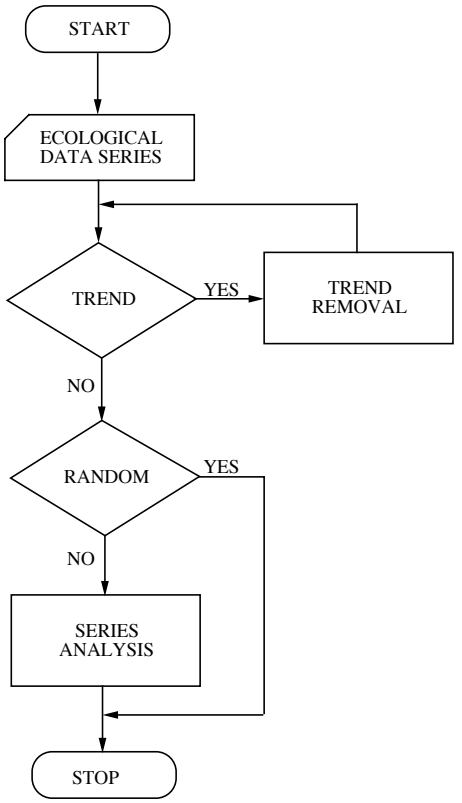


Figure 12.4 Flow diagram summarizing the steps involved in the analysis of data series.

In addition to the numerical methods discussed in the following sections, ecologists may find it useful to have a preliminary look at the data series, using the techniques of exploratory data analysis described by Tukey (1977, his Chapters 7 and 8). These are based on simple arithmetic and easy-to-draw graphs, and they may help decide which numerical treatments would be best suited for analysing the series. Exploratory data analysis for time series is also described in Chapter 14 of Venables & Ripley (2002) and in Chapter 2 of Shumway & Stoffer (2011).

12.2 Trend extraction and numerical filters

When there is a trend in a series (which is not always the case), it must be extracted from the data prior to further numerical analyses. As explained in the previous section, this is because most methods of analysis require that the series be *stationary*.

When the trend itself is of interest, it can be analysed in ecological terms (Objective 1 above). For example, Fortier *et al*. (1978) interpreted a cyclical trend in temporal changes of estuarine phytoplankton in terms of physical oceanographic forcing. In a long-term monitoring study of bacteria of sanitary importance at a lake beach, St-Louis & Legendre (1982) interpreted the significant negative slope of a water quality index computed from bacterial data (363 water samples analysed over 9 years) as an indication of deterioration of the water quality. Borcard *et al*. (2004) identified a linear trend in marine zooplankton size-class data across a coastal reef lagoon in Guadeloupe (spatial series). The trend was related to increasing salinity from the coast to the outer reef, and to decreasing phytoplankton biomass, wind speed and dissolved oxygen. In the analysis of fossil diatom assemblages along a sediment core from south-western Scotland covering the past 10000 years (101 core levels, 139 species), Legendre & Birks (2012) identified a significant temporal trend and related it to changes in the relative abundances of eight diatom species that were highly correlated with their positions along the core.

When the study goes beyond the identification of a trend (Objectives 2 *et seq*.), the analysis is normally conducted on the *residual* (or *detrended*) data series. The residual (i.e. stationary) series is obtained, for each data point i along the series, by subtracting the value estimated by the trend function at position x_i from the observed value y_i:

Residuals

$$y_{\text{res}.i} = \text{residual of } y_i = \text{observed value } (y_i) - \text{value of the trend at } x_i \qquad \textbf{(12.1)}$$

Detrending

There are cases where several trends of different natures must be extracted successively before reaching stationarity. However, because each trend extraction distorts the residuals, one must proceed with caution with detrending. The success of trend extraction may be assessed by plotting and examining the resulting trend (Objective 1) or the detrended series.

Moving averages

Window

The method of *moving averages* is often used to estimate trends, e.g. in climate-change related studies. One calculates successive arithmetic averages over $2m + 1$ contiguous data as one moves along the data series. The interval $(2m + 1)$ over which a moving average is computed is called *window*. For example, with $m = 2$, the first moving average $\bar{y}_3$ is computed over the first 5 values y_1 to y_5, the second moving average $\bar{y}_4$ is calculated over values y_2 to y_6, the third one ($\bar{y}_5$) is the average of values y_3 to y_7, and so forth. Each average value is positioned at the centre of its window. For a series of n observations, there are $(n - 2m)$ moving averages:

x_1	x_2	x_3	x_4	…	x_{n-2}	x_{n-1}	x_n
y_1	y_2	y_3	y_4	…	y_{n-2}	y_{n-1}	y_n

moving averages

$$\bar{y}_3 = \frac{1}{5}\sum_{h=1}^{5} y_h \qquad \bar{y}_4 = \frac{1}{5}\sum_{h=2}^{6} y_h \qquad \cdots \qquad \bar{y}_{n-2} = \frac{1}{5}\sum_{h=n-4}^{n} y_h$$

The general formula for moving averages is thus:

$$\bar{y}_i = \frac{1}{2m+1} \sum_{h=-m}^{m} y_{(i+h)} \qquad \textbf{(12.2)}$$

The h values corresponding to the above example, where $m = 2$, would be: –2, –1, 0, +1, and +2, respectively.

Moving averages may also be *weighted*. In such a case, each of the $2m + 1$ values within the window is multiplied by a weight w_h. Usually, values closer to the centre of the window receive larger weights. The general formula for the weighted moving average corresponding to any position (or object) x_i is:

$$\bar{y}_i = \sum_{h=-m}^{m} y_{(i+h)} w_h \Big/ \sum_{h=-m}^{m} w_h \qquad \textbf{(12.3)}$$

Choosing values for the weights depends on the underlying hypothesis. Kendall & Ord (1990, p. 3) give coefficients to be used under hypotheses of polynomial trend of the second, third, fourth, and fifth degrees. Another, simple method for assigning weights is that of *repeated moving averages*. After calculating a first series of non-weighted moving averages (eq. 12.2), a second series of moving averages is calculated using values from the first series. Thus calculation of three successive series of non-weighted moving averages produces the following results ($\bar{y}_i$) and weights w_h (Table 12.4):

first series ($m = 1$) $\quad \bar{y}_i = \frac{1}{3} \sum_{h=-1}^{1} y_{(i+h)} w_h \qquad w_0 = 1, w_{\pm 1} = 1$

second series ($m = 2$) $\quad \bar{y}_i = \frac{1}{9} \sum_{h=-2}^{2} y_{(i+h)} w_h \qquad w_0 = 3, w_{\pm 1} = 2, w_{\pm 2} = 1$

third series ($m = 3$) $\quad \bar{y}_i = \frac{1}{27} \sum_{h=-3}^{3} y_{(i+h)} w_h \qquad w_0 = 7, w_{\pm 1} = 6, w_{\pm 2} = 3, w_{\pm 3} = 1$

It is easy to check the above values by simple calculations, as shown in Table 12.4.

When using moving averages for estimating the trend of a series, one must choose the *width of the window* (i.e. choose m) as well as the *shape* of the moving average (i.e. the degree of the polynomial or the number of iterations). These choices are not simple. They depend in part on the goal of the study, namely the ecological interpretation of the *trend* itself or the subsequent analysis of *residuals* (i.e. detrended series). To estimate a cyclic trend, for instance, it is recommended to set the window width ($2m + 1$) equal to the period of the cyclic fluctuation.

Cyclic trend

Table 12.4 Calculation of repeated moving averages. Development of the numerator for the first and second series of averages.

x_1	x_2	x_3	x_4	…	x_i	…
y_1	y_2	y_3	y_4	…	y_i	…
	$\bar{y}'_2 = y_1 + y_2 + y_3$	$\bar{y}'_3 = y_2 + y_3 + y_4$	$\bar{y}'_4 = y_3 + y_4 + y_5$	…	$\sum_{h=-1}^{1} y_{(i+h)} w_h$ $w_0 = 1, w_{\pm 1} = 1$	…
		$\bar{y}''_3 = \bar{y}'_2 + \bar{y}'_3 + \bar{y}'_4$	$\bar{y}''_4 = \bar{y}'_3 + \bar{y}'_4 + \bar{y}'_5$	…	$\sum_{h=-2}^{2} y_{(i+h)} w_h$	…
		$\bar{y}''_3 = y_1 + 2y_2 + 3y_3 + 2y_4 + y_5$	…	…	$w_0 = 3, w_{\mp 1} = 2, w_{\mp 2} = 1$	

Slutzky-Yule effect

Trend extraction by moving averages may add to the detrended series an artificial periodic component, which must be identified before analysing the series. This phenomenon is called the *Slutzky-Yule effect*, because these two statisticians independently drew attention to it in 1927. According to Kendall (1976, pp. 40-45) and Kendall *et al*. (1983, pp. 465-466), the average period of this artificial component (T) is calculated using the $(2m + 1)$ weights w_i of the moving average formula (eq. 12.3)*:

$$T = 2\pi/\theta \quad \text{for angle } \theta \text{ in radians, or} \quad T = 360°/\theta \quad \text{for angle } \theta \text{ in degrees,}$$

$$\text{where} \qquad \cos\theta = \left| \sum_{h=1}^{2m+1} (w_{h+1} - w_h)(w_h - w_{h-1}) \right| \Big/ \sum_{h=1}^{2m+2} (w_h - w_{h-1})^2 \qquad \textbf{(12.4)}$$

The values of the weights located outside the window are zero: $w_0 = 0$ and $w_{2m+2} = 0$. For example, using the weights of the second series of repeated moving averages above ($m = 2$):

$$[w_h] = [1 \;\; 2 \;\; 3 \;\; 2 \;\; 1]$$

* In Kendall (1976) and Kendall *et al*. (1983) and previous editions of *The Advanced Theory of Statistics, Vol. 3*, there is a printing error in the formula for the Slutzky-Yule effect. In the first parenthesis of the last term of their numerator, the printed sign for the second weight (w_{2m+1}) is positive; this sign should be negative, as in eq. 12.4, giving (0 – 1) in our numerical example. However, their numerical example is correct, i.e. it is computed with $-w_{2m+1}$, not $+w_{2m+1}$.

gives

$$\cos\theta = \frac{|(2-1)(1-0) + (3-2)(2-1) + (2-3)(3-2) + (1-2)(2-3) + (0-1)(1-2)|}{(1-0)^2 + (2-1)^2 + (3-2)^2 + (2-3)^2 + (1-2)^2 + (0-1)^2} = \frac{3}{6}$$

from which it follows that $\theta = 1.047 \text{ rad} = 60°$

and thus: $T = 2\pi/1.047 = 360°/60° = 6$

If, after detrending by this method of *repeated moving averages*, the analysis of the series resulted in a period $T \approx 6$, this period would probably be a by-product of the moving average procedure. It would not correspond to a component of the original data series, so that one should not attempt to interpret it in ecological terms. If a period $T \approx 6$ was hypothesized to be of ecological interest, one should use different weights for trend extraction by moving average analysis.

Analytical method

The most usual approach for estimating trends is the *analytical method*. It consists in fitting a regression model to the whole series, using the least squares approach or some other method. The matter was fully reviewed in Section 10.3. Smoothing methods such as splines and LOWESS can also be used (Subsection 10.3.8). The model for the trend may be linear, polynomial, exponential, logistic, etc. The main advantages of trend extraction based on regression are: the explicit choice of a model by the investigator, and the ease of calculation using a statistical package. The main problem is that a new regression must be calculated upon addition of one or several observations to the data series, which may generate different values for the regression coefficients. However, as the series gets longer, estimates of the regression coefficients become progressively more stable.

Variate difference

Contrary to the above methods, where the estimated trend was subtracted from the observed data (eq. 12.1), the *variate difference method* directly detrends the series. It consists in replacing each value y_i by the difference $(y_{i+1} - y_i)$. As in the case of repeated moving averages, differences may be calculated not only on the original data, but also on data resulting from previous detrending. If this is repeated on progressively more and more detrended series, the variance of the series usually stabilizes rapidly. The variate difference method, when applied once or a few times to a series, can successfully remove any polynomial trend. Only exponential or cyclic trends may

Cyclic trend

sometimes resist the treatment. The method may be used to remove any cyclic trend whose period T is known, by using differences $(y_{i+T} - y_i)$; however, this is fully successful only in cases where T is an integer multiple of the sampling interval Δ. One must remember that this method does not model the trend that is removed from the data series as a data vector; hence, the trend cannot be studied independently.

Filtration

In some instances, ecologists may also wish to eliminate the random *noise* component from the data series, in order to better evidence the ecological phenomenon under study. This operation, whose aim is to remove high-frequency variability from the series, is called *filtration*. In a sense, filtration is the complement of trend

Filter

extraction, since the former removes high-frequency components of the series and the latter, low-frequency components. Specialists of series analysis often use the term *filter* for any preliminary treatment of the series, whether the extraction of low frequencies (trend) or the removal of high frequencies (noise). Within the context of spectral analysis (Section 12.5), filtration of the series is often called "prewhitening". This refers to the fact that filtration flattens the spectrum of a series and makes it similar to the spectrum of white light. The reciprocal operation (called "recolouring") fits the spectrum (calculated on the filtered series) in such a way as to make it representative of the nonfiltered series. The sequence of operations — prewhitening of the series, followed by computation of the spectrum on the filtered series, and finally recolouring of the resulting spectrum — finds its justification in the fact that spectra that are more flat are also more precisely estimated.

In addition to filters, which aim at extracting low frequencies (trends), computer programs for series analysis offer a variety of numerical filters that allow the removal, or at least the reduction, of any component located outside a given frequency band (passband). It is thus possible, depending on the objective of the study, to select the high or low frequencies, or else a band of intermediate frequencies. It is also possible to eliminate a band of intermediate frequencies, which is the converse of the latter filter. Generally, these numerical filters are found in programs for spectral analysis (Section 12.5), but they may also be used to filter series prior to analyses using the methods described in Sections 12.3 and 12.4. In most cases, filtering data series (including trend extraction) requires solid knowledge of the techniques, because filtration always distorts the original series and thus influences further calculations. It is therefore better to do it under the supervision of an experienced colleague.

12.3 Periodic variability: correlogram

The systematic component of a stationary series is called *periodic variability*. There are several methods available for analysing this type of variability. Those discussed in the present section, namely the autocovariance and autocorrelation (serial correlation) and the cross-covariance and cross-correlation, are all extensions, to the analysis of data series, of statistical methods described in earlier chapters. These methods of analysis have been extensively used in ecology.

At this stage of series analysis, it is assumed that the data series is *stationary*, either because it originally exhibited no trend or as the result of detrending (Section 12.2). It is also assumed that variability is large enough to emerge from random noise.

A general approach for analysing periodic variability is derived from the concepts of covariance and correlation defined in Chapter 4. The methods are called *autocovariance* and *autocorrelation analysis*. The approach is to quantify the relationships between successive terms of the data series. These relationships reflect the pattern of periodic variability.

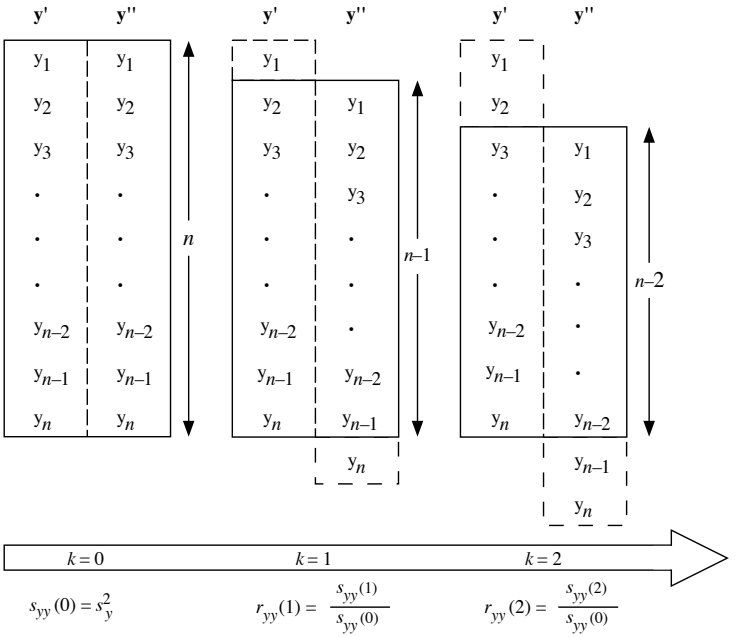


Figure 12.5 Calculation of autocovariance (s_{yy}) and autocorrelation (r_{yy}). Stepwise shift of a data series relative to itself, with successive lags of k units. The number of terms involved in the calculation ($n - k$) decreases as k increases.

1 — Autocovariance and autocorrelation

Autocovariance measures the covariance of the series with itself, computed as the series is progressively shifted with respect to itself (Fig. 12.5). Because second-order stationarity is assumed in the calculation of autocovariance and autocorrelation (Section 12.1), all coefficients will be computed using the same mean and variance, estimated from the whole series, even though individual coefficients involve only part of the data. The overall mean is $\bar{y}$. For the common variance, the sum of squared deviations from $\bar{y}$ is divided by n instead of $(n - 1)$, as in Moran's I coefficient of spatial correlation (eq. 13.1); this is the maximum-likelihood estimator of the variance:

$$s_{yy}(0) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \textbf{(12.5)}$$

This is the covariance of the series with itself when there is no shift. The notation $s_{yy}(0)$ indicates a lag of zero, or lag $k = 0$.

When the series is shifted relative to itself by one unit (lag $k = 1$), the left-hand copy of the series in Fig. 12.5 loses observation y_1 and the right-hand copy loses observation y_n. The two truncated series, each of length $(n - 1)$, are compared. For a lag of k units, the covariance $s_{yy}(k)$ is computed from the $(n - k)$ terms remaining in the two truncated series, using the mean and n value from the whole series to insure that the covariances remain comparable:

Auto-covariance

$$s_{yy}(k) = \frac{1}{n} \sum_{i=1}^{n-k} (y_{i+k} - \bar{y})\,(y_i - \bar{y}) \qquad \textbf{(12.6)}$$

That equation is similar to that of the covariance (eq. 4.4). In correlograms (below), the autocovariance is estimated for several successive lags k. In specific applications, researchers may decide on biological grounds how long the lag should be to compute the autocovariance of the variable under study.

In eq. 4.7, the Pearson coefficient of linear correlation between variables y_j and y_k is computed by dividing their covariance by the product of their standard deviations:

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

In a similar way, the *autocorrelation* of a series $r_{yy}(k)$ is computed as the ratio of its autocovariance $s_{yy}(k)$ (eq. 12.6) to its variance $s_{yy}(0)$ (eq. 12.5):

Auto-correlation

$$r_{yy}(k) = \frac{s_{yy}(k)}{s_{yy}(0)} \qquad \textbf{(12.7)}$$

Equation 12.6 is a good estimator of autocorrelation when $(n - k)$ is reasonably large. The autocorrelation is also called *serial correlation*. It measures the average dependence of the values in the series on values found at a distance of k lags.

One may be tempted to compute $r_{yy}(k)$ using the Pearson linear correlation formula (eq. 4.7) between terms y_i and y_{i+k} of the series, for the $n - k$ pairs of corresponding values in the observed and shifted series (Fig. 12.5). This is not recommended, however, because the mean and variance estimates used for computing eq. 4.7 change with lag k, so that $r_{yy}(k)$ would not produce a set of comparable autocorrelation coefficients (Jenkins & Watts, 1968; Venables & Ripley, 2002).

Since the number of terms $(n - k)$ involved in the calculation of the autocovariance or autocorrelation decreases as k increases, it follows that, as k increases, the precision of the estimate, the number of degrees of freedom available, and consequently the power of the test of significance decrease. The largest interpretable lag is often considered to be about $k_{max} = n/3$; Venables & Ripley (2002) use $10 \log_{10}(n)$ as the default value for the largest lag in function ***acf()*** in R. Table 12.5 gives the values of autocovariance and autocorrelation for the artificial stationary series of Fig. 12.2b.

Table 12.5 Autocovariance and autocorrelation coefficients (eq. 12.7) for the artificial series of Fig. 12.2b, after detrending (i.e. periodic signal + noise components only). For each successive lag, the series is shifted by one sampling interval. Values corresponding to odd lags are not shown. The autocovariance and autocorrelation coefficients are plotted against lag in Fig. 12.6.

Lag	Autocovariance $s_{yy}(k)$	Autocorrelation $r_{yy}(k)$
0	3.07	1.00
2	1.01	0.33
4	–0.90	–0.29
6	–0.28	–0.09
8	–0.29	–0.10
l0	–0.87	–0.28
12	–0.36	–0.12
14	–0.27	–0.09
16	–0.64	–0.21
18	0.33	0.11
20	1.22	0.40
22	0.47	0.15

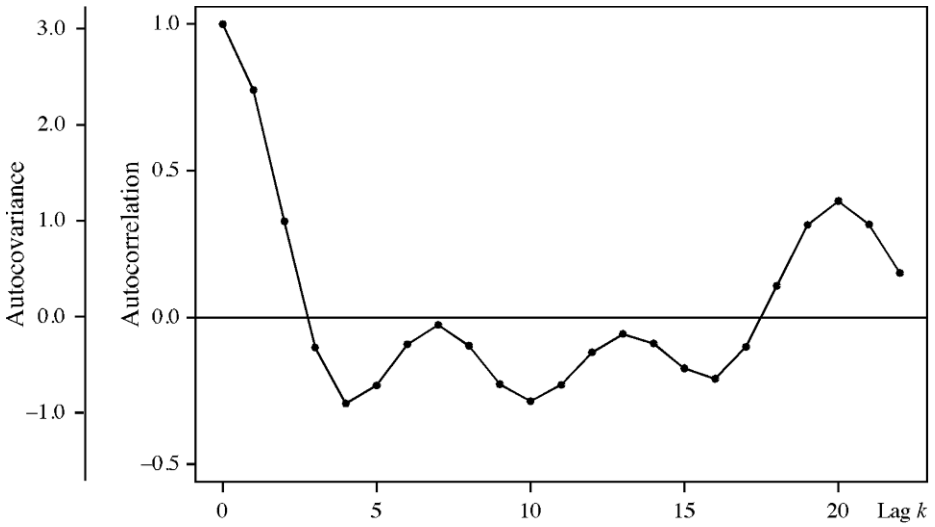


Figure 12.6 Correlogram (autocovariance and autocorrelation; values from Table 12.5) for the artificial series of Fig. 12.2b, after detrending (i.e. periodic signal + noise components only).

Autocorrelogram

The autocorrelation (or autocovariance) coefficients are plotted as a function of lag k (abscissa), in a graph called *autocorrelogram* (or *correlogram*, for simplicity). Autocorrelation coefficients range between +1 and –1. The scale factor between the autocorrelation and autocovariance coefficients is the variance of the series (eq. 12.5). In Fig. 12.6, this factor is $s_{yy}(0) = 3.07$; it is shown in Table 12.5 at lag $k = 0$.

The interpretation of correlograms is based on the following reasoning. At lag $k = 0$, the two copies of the series (**y**' and **y**") have the exact same values facing each other (Fig. 12.5), hence $r_{yy}(0) = +1$. With increasing lag k, corresponding values in the series **y**' and **y**" move farther apart and $r_{yy}(k)$ decreases. This is what is happening, in the numerical example, for lags up to $k = 4$ (Table 12.5 and Fig. 12.6). In series where periodic variability is present (with period T_p), increasing k eventually brings similar values to face each other again (at lag $k = T_p$), with peaks facing peaks and troughs facing troughs, hence a high positive value of $r_{yy}(k)$. The value of $r_{yy}(k = T_p)$ is always smaller than 1, however, because there is always noise in data and because natural periodic phenomena seldom repeat themselves perfectly. Negative autocorrelation often reaches its maximum at $k = T_p/2$ because the signals in **y**' and **y**" are then maximally out of phase.

A practical problem occurs when there are several periodic signals in a series; this may increase the complexity of the correlogram. Nevertheless, high positive values in a correlogram may generally be interpreted as indicative of the presence of periodic variability in the series. For the numerical example, Fig. 12.6 indicates that there is a major periodicity at $k = 20$, corresponding to period $T = 20$; this interpretation is supported by the low value of $r_{yy}(10)$. Period $T = 20$ is indeed the distance between corresponding peaks or troughs in the series of Fig. 12.2b. Other features of the correlogram may be indicative of additional periods (which is the case here, as can be seen by examining Fig. 12.2b) or may simply be the result of random noise.

Confidence intervals can be computed and drawn on a correlogram to identify the values that are significantly different from zero. The confidence interval is usually represented on the correlogram as a two-standard-error band. If the data can be assumed to be normal, independent (in the sense of *not autocorrelated*, Box 1.1) and identically distributed, the confidence interval of r_{yy} can be computed through the usual formula for confidence intervals of correlation coefficients. In most time series analyses, however, there is an assumption that the data are autocorrelated. It is thus more appropriate to compute confidence intervals under a moving average (MA) model (eq. 12.31) (Venables & Ripley, 2002). Both methods of calculation are available in the R function ***plot.acf()*** (Section 12.8).

Harmonic

When the series is long, its correlogram may exhibit significant values for *harmonics* (integer multiples) of the period present in the signal (T_{series}). This is a normal phenomenon, which is generally not indicative of additional periodicity in the data series. However, when a value of the correlogram statistic is noticeably larger for a harmonic period than for the basic period, one can conclude that the harmonic is also a true period of the series.

For short series, autocorrelograms should only be computed when the series include very strong periodic components. This is because the test of significance is not very powerful, i.e. the probability of rejecting the null hypothesis of no autocorrelation is small when a periodic component is present in short series. When there is *more than one periodic component* in a series, correlograms should generally not be used, even with long series, because components of different periods may interfere with one another and prevent the correlogram from showing significance (see also the next paragraph). Periodograms (Section 12.4) should be used instead. Finally, when the data are *not equispaced* and one does not wish to interpolate, methods developed for *spatial correlation* analysis, which do not require equal spacing of the data, may be used (Section 13.1). Special forms of spatial correlation coefficients allow the analysis of series of *qualitative* data (last paragraph of Subsection 13.1.1).

It may happen that periods present in the series do not appear in a correlogram, because they are concealed by other periods accounting for larger fractions of the variance of the series. When one or several periods have been identified using a first correlogram, one may remove these periods from the series using one of the methods recommended in Section 12.2 for cyclic trends and compute a new correlogram for the detrended series. It could bring out previously concealed periods. This is not without risk, however, because successively extracting trends rapidly distorts the residuals. Approaches better adapted to series containing multiple periods are discussed in Sections 12.4 and 12.5.

The following numerical example and ecological applications illustrate the computation and use of correlograms.

Numerical example. Consider the following series of 16 data points (quantitative variable):

$$2, 2, 4, 7, 10, 5, 2, 5, 8, 4, 1, 2, 5, 9, 6, 3$$

Table 12.6 illustrates the computation of the autocorrelation coefficients. These could be plotted as a function of lag (k) to form a correlogram, as in Figs. 12.6 and 12.7b. The coefficients clearly point to a dominant period at $k = 5$, for which autocorrelation is positive and maximum. This approximately corresponds to the average distance separating successive maximum values, as well as successive minima, along the data series.

Ecological application 12.3a

In order to study the spatial variability of coastal marine phytoplankton, Platt *et al*. (1970) measured chlorophyll *a* along a transect 8 nautical miles long, at 10 m depth and intervals of 0.1 naut. mi. (1 naut. mi. = 1852 m). The resulting 80 values are shown in Fig. 12.7a.

The series exhibited a clear linear *trend*, which was extracted at the beginning of the analysis. Autocorrelation coefficients were computed from the residual series, up to lag $k = 10$, because the series was quite short (Fig. 12.7b). The position of the first *zero* in the *correlogram* was taken as indicative of the average apparent *radius* of phytoplankton patches along the transect. The model underlying this interpretation is that of circular patches, separated by average distances equal to their average diameter. In such a case, it is expected that the second

Table 12.6 Computation of the autocorrelation coefficients for the data of the numerical example. Boxes delimit the values included in each calculation. Note how the highest values are facing each other at lag 5, where the autocorrelation coefficient is maximum.

Lag	Data series	Autocorrelation $r_{yy}(k)$
k=0	2 2 4 7 10 5 2 5 8 4 1 2 5 9 6 3 2 2 4 7 10 5 2 5 8 4 1 2 5 9 6 3	1.000
k=1	2 | 2 2 4 7 10 5 2 5 8 4 1 2 5 9 6 3 2 2 4 7 10 5 2 5 8 4 1 2 5 9 6 | 3	0.313
k=2	2 2 | 4 7 10 5 2 5 8 4 1 2 5 9 6 3 2 2 4 7 10 5 2 5 8 4 1 2 5 9 | 6 3	–0.544
k=3	2 2 4 | 7 10 5 2 5 8 4 1 2 5 9 6 3 2 2 4 7 10 5 2 5 8 4 1 2 5 | 9 …	–0.472
k=4	2 2 4 7 | 10 5 2 5 8 4 1 2 5 9 6 3 2 2 4 7 10 5 2 5 8 4 1 2 | 5 …	0.105
k=5	2 2 4 7 10 | 5 2 5 8 4 1 2 5 9 6 3 2 2 4 7 10 5 2 5 8 4 1 | 2 …	0.323
k=6	2 2 4 7 10 5 | 2 5 8 4 1 2 5 9 6 3 2 2 4 7 10 5 2 5 8 4 | 1 …	–0.107
etc.	etc.	etc.

zero would occur at a lag three times that of the first zero, as was indeed observed on the correlogram. In the present case, the average *diameter* of phytoplankton patches and the distance separating them appeared to be ca. 0.5 naut. mi.

Ecological application 12.3b

Steven & Glombitza (1972) sampled tropical phytoplankton and chlorophyll at a site off Barbados during nearly three years. Sampling was approximately fortnightly. The physical environment there is considered to be very stable over the year. The most abundant phytoplankton species, in surface waters, is the filamentous cyanobacterium *Trichodesmium thiebaudii*. Data were concentrations of chlorophyll *a* and of *Trichodesmium* filaments.

The raw data were subjected to two transformations: (1) computation of *equispaced* data at 15-day intervals by interpolation, and (2) *filtration* intended to reduce the importance of non-dominant variations. The filtered data are shown in Fig. 12.8a, where the synchronous variations of the two variables are obvious. Correlograms for the nonfiltered (Fig. 12.8b) and filtered (Fig. 12.8c) series clearly show the same periodic signal, of ca. 8 lags × (15 days lag^{-1}) = 120 days. Nonfiltered data provide the same information as the filtered series, but not quite as clearly. According to the authors, these periodic variations could be an example of free

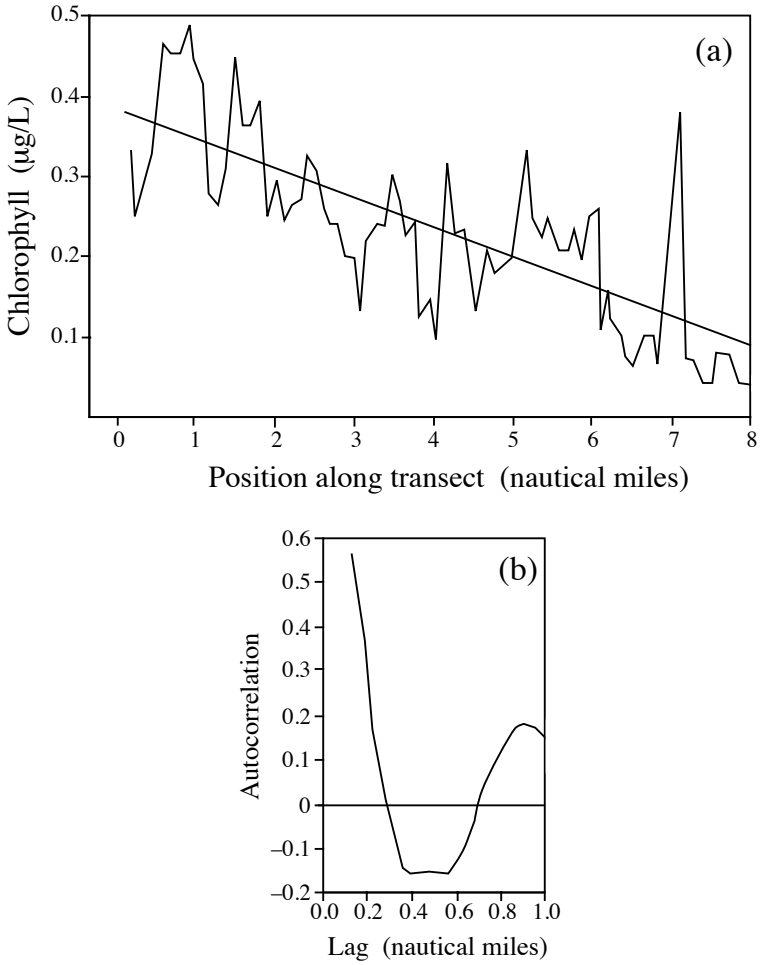


Figure 12.7 Chlorophyll *a* concentrations in a coastal marine environment, along a transect 8 naut. miles long in St. Margaret's Bay (Nova Scotia, Canada). (a) Data series exhibiting a linear trend, and (b) correlogram of the detrended series where lags (abscissa) are given as distances along the transect. After Platt *et al*. (1970).

oscillations, since they seemed independent of any control by the environment, which was stable the year round. The same ecological application will be used again below to illustrate the calculation of cross-correlation (next subsection) and Schuster's periodogram (Section 12.4).

Wilson & Dawe (2006) used autocorrelograms to compare variations in population densities of marine foraminifera with monsoonal rainfall data. Dutilleul (2011, his Sections 6.2.1 and 6.3.2) discussed applications of autocorrelation to several data

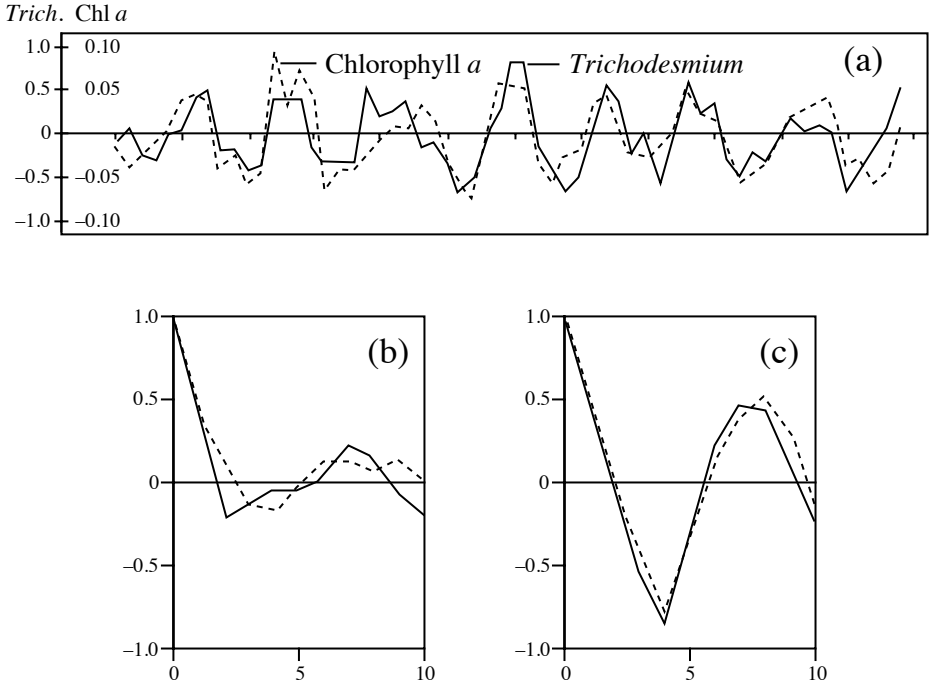


Figure 12.8 (a) Filtered time series of chlorophyll *a* and *Trichodesmium* in tropical surface waters off Barbados. Marks along the abscissa are spaced by 75 days. On the ordinate, units are 10^3 filaments *Trichodesmium* L^{-1} and µg chlorophyll *a* L^{-1}. Correlograms of (b) the nonfiltered series and (c) the filtered series. After Steven & Glombitza (1972).

series: maternal behaviour of the Wistar rat (a strain of albino rats) in the laboratory, observed every two hours during five days (his Fig. 6.7, b1 and b2); yearly mean sunspot numbers for the period 1749-1994 (his Fig. 6.8, b and c); monthly atmospheric CO_2 concentrations at Mona Laua, Hawaii, from 1965 through 2004 (his Fig. 6.9, c and d); daily mean temperatures in air and soil in the Gault Nature Reserve (Québec) over thirty days in June 2004 (his Fig. 6.3, b-c and f-g); and hourly mean temperatures in air and soil in the same nature reserve over eight days in June (his Fig. 6.10, c-d and g-h).

2 — Cross-covariance and cross-correlation

In order to determine the extent to which two data series exhibit concordant periodic variations, a method closely related to autocovariance and autocorrelation can be used. This method has two variants called *cross-covariance* and *cross-correlation* (or *lag correlation*).

Consider two series, $\mathbf{y}_j$ and $\mathbf{y}_l$, of the same length. One is progressively shifted with respect to the other, with lags $k = 1, 2, \ldots$ As the lag increases, the zone of overlap between the two series shortens. *Cross-covariance* of order k is computed in a way analogous to autocovariance. As in eq. 12.6 for autocovariance, the means $\bar{y}_j$ and $\bar{y}_l$ of the full series are used to compute the cross-covariance $s_{jl}(k)$ between the two series for lag k:

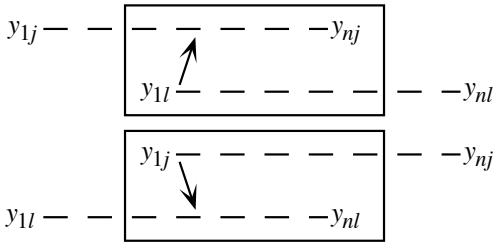
Cross-covariance

$$s_{jl}(k) = \frac{1}{n} \sum_{i=1}^{n-k} [y_{(i+k)j} - \bar{y}_j] [y_{il} - \bar{y}_l] \qquad \textbf{(12.8)}$$

When $k = 0$ (no shift), eq. 12.6 becomes the maximum likelihood estimator of the covariance between the variables:

$$s_{jl}(0) = \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j) (y_{il} - \bar{y}_l)$$

Equation 12.8 shows an important difference between *cross-covariance* and *autocovariance*, namely that the relative direction in which a series is shifted with respect to the other must be taken into account. Indeed, shifting series $\mathbf{y}_j$ "to the right" by k units with respect to series $\mathbf{y}_l$ is not equivalent to shifting it "to the left" because the direction of the implied causal relationship (arrows in the figure) is not the same:



The value of cross-covariance for lag k would be different if $\mathbf{y}_j$ and $\mathbf{y}_l$ were interchanged in eq. 12.8; in other words, generally $s_{jl}(k) \neq s_{lj}(k)$. In order to distinguish between the two sets of cross-covariances, one set of shifts is labelled as positive and the other as negative. The choice of the positive and negative directions is arbitrary and without consequence. In eq. 12.8, if the cross-covariance of $\mathbf{y}_j$ relative to $\mathbf{y}_l$ is identified as $s_{jl}(k)$, the converse would be labelled $s_{jl}(-k)$. No distinction was made between the two relative shift directions in autocovariance (eq. 12.6) because $s_{yy}(+k) = s_{yy}(-k)$. When the direction of the causal relationship is known, there is no need to compute cross-covariance for both positive and negative shifts, although computer functions may automatically compute them, e.g. function ***ccf()*** in R.

The cross-covariance is generally plotted as a function of the positive and negative lags k, to the right and to the left of $k = 0$. The alternative is to plot the two sets on the

positive side of the abscissa using two different symbols. Maximum cross-covariance does not necessarily occur at $k = 0$. Sometimes, the dependence between the two series is maximum at a lag $k \neq 0$. In predator-prey interactions for example, cross-covariance may be maximum for a lag corresponding to the response time of the predator population (*target variable*) to changes in the number of prey (*predictor variable*). One then says that the target variable *lags* the causal variable.

Cross-covariance can be transformed into *cross-correlation*. To do so, the cross-covariance $s_{jl}(k)$ is divided by the product of the corresponding standard deviations, which are the square roots of the variance $ss_{jj}(0)$ and $ss_{ll}(0)$ (eq. 12.5):

Cross-correlation

$$r_{jl}(k) = \frac{s_{jl}(k)}{\sqrt{s_{jj}(0)\, s_{ll}(0)}} \qquad \textbf{(12.9)}$$

Cross-correlogram

As for cross-covariance, cross-correlation is defined for $+k$ and $-k$. Values are plotted as a function of k in a *cross-correlogram*. Fortier & Legendre (1979) used Kendall's τ (Section 5.3) instead of Pearson's r for computing cross-correlations between series of quantitative variables which were *not linearly related*. They called this measure *Kendall's cross-correlation*. It may also be applied to series of *semiquantitative* data; Spearman's r (Section 5.3) could be used instead of Kendall's τ. Legendre & Legendre (1982) proposed to extend this approach to *qualitative* data under the name *cross-contingency*. In that case, contingency statistics (X^2 or uncertainty coefficients; Section 6.2) are computed for the two series as one is progressively shifted with respect to the other.

When several ecological variables are observed simultaneously, the resulting *multidimensional series* may be analysed using cross-covariance or cross-correlation. Such methods are obviously of interest in ecology, where variation in one variable is often interpreted in terms of variation in others. However, eq. 12.9 considers only two series at a time; for multidimensional data series, it is sometimes useful to extend the concept of *partial correlation* (Sections 4.5 and 5.3) to the approach of cross-correlation. In Ecological application 12.3d, Fréchette & Legendre (1982) used Kendall's partial (partial τ; Section 5.3) cross-correlation to analyse an ecological situation involving three variables.

Ecological application 12.3c

In their study on temporal variability of tropical phytoplankton (Ecological application 12.3b), Steven & Glombitza (1972) compared the variations in concentrations of chlorophyll *a* and *Trichodesmium*, using cross-correlations (Fig. 12.9). The cross-correlogram shows that changes in the two variables were in phase, with a period of 8 lags × 15 days lag^{-1} = 120 days. Filtration of the data series brought but a small improvement to the cross-correlation. These results confirm the conclusions drawn from the correlograms (Fig. 12.8), and show that variations of chlorophyll *a* concentration, in surface waters, were due to changes in the concentration of *Trichodesmium* filaments. This same application will be further discussed in Section 12.4 (Ecological application 12.4e).

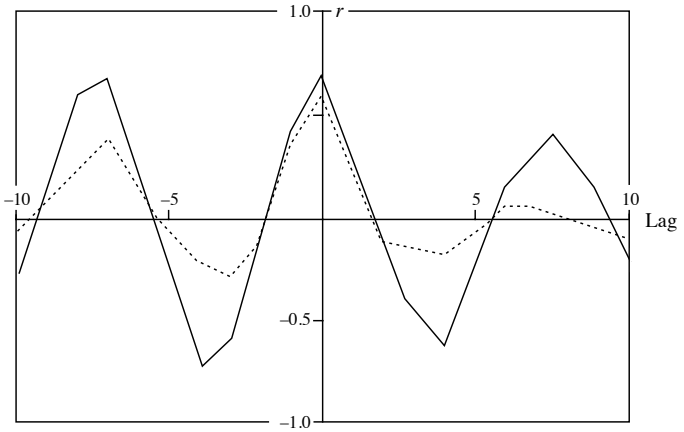


Figure 12.9 Cross-correlations between temporal changes in concentrations of chlorophyll *a* and *Trichodesmium*, in tropical surface waters, computed on nonfiltered (solid line) and filtered (dotted line) data series. After Steven & Glombitza (1972).

Ecological application 12.3d

At an anchor station in the St. Lawrence Estuary (Québec), Fréchette & Legendre (1982) determined the photosynthetic capacity of phytoplankton (P^B_{max}) hourly, during six consecutive days. The sampling area was subjected to internal tides, which drove changes in two important physical variables: (1) vertical oscillations of the water mass (characterized in this study by the depth of isopycnal $\sigma_t = 22$, i.e. the depth where the density of water was 1022 kg m^{-3}), and (2) variations in the vertical stability of the upper water column, estimated as the density gradient between 1 and 25 m. Two hypotheses could explain the observed effects of internal tides on P^B_{max}: (1) upwelling under the effect of incoming internal tides, up to the depths where sampling took place, of deeper water containing phytoplankton with lower P^B_{max}, or (2) adaptation of P^B_{max} to changes in the vertical stability of the upper water column.

Since the two physical variables were controlled by the same mechanism (i.e. internal tides), it was not easy to identify their specific contributions to phytoplankton photosynthesis. This was achieved by computing two Kendall's partial cross-correlations (partial τ): (1) between P^B_{max} and the depth of $\sigma_t = 22$, controlling for the effect of vertical stability, and (2) between P^B_{max} and stratification, controlling for vertical displacement. When calculating the partial cross-correlations, the response variable (P^B_{max}) was shifted relative to the two potentially causal (physical) variables until a maximum value was reached.

The authors concluded that the photosynthetic activity of phytoplankton responded to changes in the vertical stability of the water column, driven by internal tides. This was interpreted as an adaptation of the cells to periodic variations in their light environment.

Another example of cross-correlation applied to an ecological data series can be found in Wilson & Dawe (2006), who used cross-correlograms to compare variations in population densities of marine foraminifera with monsoonal rainfall data.

For series with irregular lag or missing data, the multivariate variogram (Subsection 13.1.4) can be used to detect periodic phenomena in univariate or multivariate quantitative data series. Likewise, the Mantel correlogram (Subsection 13.1.6) can be used to detect periodic phenomena in irregular univariate quantitative, semiquantitative or qualitative data series, and in multivariate series involving variables of any precision level. This type of correlogram is computed from a distance matrix among the observations in the series.

12.4 Periodic variability: periodogram

In addition to the relatively simple methods discussed in the previous section, there is another general approach to the study of periodic variability, called *harmonic analysis*. This approach is mathematically more complex than correlogram analysis, but it is often better adapted to the study of ecological data series. Results of harmonic analysis are generally plotted in a graph called *periodogram*.

1 — Periodogram of Whittaker and Robinson

The simplest way to approach harmonic analysis is to examine a *Buys-Ballot table*. Assume that a series of n quantitative observations is characterized by a period T_{series}. If $T = T_{\text{series}}$ is known, the series can be split into n/T sequences, each containing T observations. A Buys-Ballot table (Table 12.7) is a double-entry table whose rows contain the $r = n/T$ sequences of T observations. The number of columns corresponds to the known or assumed period of the data series. If $T = T_{\text{series}}$, the r successive rows in the table are repetitions of the same oscillation, although the actual values in any column (j) are generally not identical because of noise. Calculating the mean value for each column ($\bar{y}_{T,\,j}$) and comparing these means is a way of characterizing the variation within period T_{series}.

When there exists a hypothesis concerning the value of T_{series} (e.g. a diurnal cycle), Buys-Ballot tables may be constructed for this value and also for neighbouring lower and higher values T_k. As the period of the table (T_k) approaches that of the series (T_{series}), values within each column become more similar, so that all maximum values tend to be located in one column and all minimum values in another. As a result, the difference between the highest and lowest mean values is maximum when period T_k of the table is the same as period T_{series} of the series. The *amplitude* of a Buys-Ballot

Amplitude

Table 12.7 Buys-Ballot table. Allocation of data from a series containing n observations to the rows of the table.

	1	2	3	...	T
1	y_1	y_2	y_3	...	y_T
2	y_{T+1}	y_{T+2}	y_{T+3}	...	y_{2T}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
r	$y_{(r-1)T+1}$	$y_{(r-1)T+2}$	$y_{(r-1)T+3}$	...	$y_{rT} = y_n$
$\bar{y}_T$	$\bar{y}_{T,1}$	$\bar{y}_{T,2}$	$\bar{y}_{T,3}$	...	$\bar{y}_{T,T}$

table is some measure of the variation found among the columns of the table. It may be measured by the *range* of the column means (Whittaker & Robinson, 1924):

Range

$$[\bar{y}_{\max} - \bar{y}_{\min}] \qquad \textbf{(12.10)}$$

or by the *standard deviation* of the column means (Enright, 1965):

Standard deviation

$$\sqrt{\frac{1}{T_k}\sum_{j=1}^{T_k}(\bar{y}_{T_k,j} - \bar{y}_{T_k})^2}, \quad \text{where} \quad \bar{y}_{T_k} = \frac{1}{T_k}\sum_{j=1}^{T_k}\bar{y}_{T_k,j} \qquad \textbf{(12.11)}$$

When the period T of interest is not an integer multiple of the interval between two observations, a problem occurs in the construction of the Buys-Ballot table. The solution proposed by Enright (1965) is to construct the table with a number of columns equal to the largest integer that is less than or equal to the period of interest, T. Observations are attributed to the columns in sequence, as usual, leaving out an observation here and there in such a way that the average rate of advance in the series, from row to row of the Buys-Ballot table, is T. This is done, formally, by using the following formula for the mean of each column j:

$$\bar{y}_{T,j} = \frac{1}{r}\sum_{i=1}^{r} y_{[(i-1)T+j]} \qquad \textbf{(12.12)}$$

where r is the number of rows with data in column j of the table. The subscript of y is systematically rounded up to the next integer. Thus, for example, if $T = 24.5$, $\bar{y}_{T,1}$ is estimated from values $\{y_1, y_{26}, y_{50}, y_{75}, y_{99}, y_{124}, \text{etc.}\}$ found in rows $i = \{1, 2, 3, 4, 5, 6, \text{etc.}\}$ of the table;

in other words, intervals of 24 and 25 units are successively used, to give an average period $T = 24.5$. This modified formula is required to understand Ecological application 12.4a, where fractional periods are used.

When studying an empirical data series, the period T_{series} is not known *a priori*. Even when some hypothesis is available concerning the value of T_{series}, one may want to check whether the hypothesized value is the one that best emerges when analysing the data. In both situations, estimating T_{series} becomes the purpose of the analysis. The values of amplitude, computed for different periods T, may be plotted together as a periodogram in order to determine which period best characterizes the data series.

Periodogram

The *periodogram of Whittaker & Robinson* is a graph in which the measures of amplitude (eq. 12.10 or 12.11) are plotted as a function of periods T_k. According to Enright (1965), periodograms based on the statistic of eq. 12.11 are more internally consistent than those based on eq. 12.10. Various ways have been proposed for testing the significance of statistic 12.11 (reviewed by Sokolove & Bushell, 1978); these tests are only asymptotically valid, so that they are not adequate for short time series.

Numerical example. Consider again the series (2, 2, 4, 7, 10, 5, 2, 5, 8, 4, 1, 2, 5, 9, 6, 3) used in Subsection 12.3.1 to compute Table 12.6. In order to examine period $T_k = 4$, for instance, the series is cut into segments of length 4 as follows:

$$2, 2, 4, 7; \quad 10, 5, 2, 5; \quad 8, 4, 1, 2; \quad 5, 9, 6, 3$$

and distributed in the successive rows of the table. Buys-Ballot tables for periods $T_k = 4$ and 5 are constructed as follows:

$T = 4$	1	2	3	4
Row 1	2	2	4	7
Row 2	10	5	2	5
Row 3	8	4	1	2
Row 4	5	9	6	3
Means	6.25	5	3.25	4.25

Range = 3, standard deviation = 1.0951

$T = 5$	1	2	3	4	5
Row 1	2	2	4	7	10
Row 2	5	2	5	8	4
Row 3	1	2	5	9	6
Row 4	3				
Means	2.75	2	4.67	8	6.67

Range = 6, standard deviation = 2.2708

The range is calculated using eq. 12.10 and the standard deviation with eq. 12.11. Repeating the calculations for $k = 2$ to 8 produces the periodogram in Fig. 12.10.

Interpretation of the periodogram may be quite simple. If one and only one oscillation is present in the series, the period with maximum amplitude is taken as the best estimate for the true period of this oscillation. Calculation of the periodogram is made under the assumption that there is *a single stable period* in the series. If several periods are present, the periodogram may be so distorted that its interpretation could lead to erroneous conclusions. Enright (1965) provides examples of such distortions, using artificial series. Other methods, discussed below, are better adapted to series with several periods.

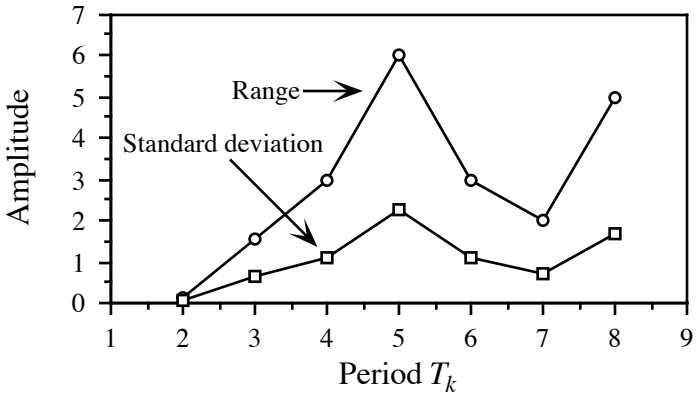


Figure 12.10 Periodogram of Whittaker and Robinson for the artificial data series. The amplitude statistics plotted in the periodogram may be either the range or the standard deviation of the column means in the Buys-Ballot tables.

Ecological application 12.4a

Enright (1965) re-examined 17 time series taken from the literature, which described the activity of animals as diverse as the chaffinch (bird), laboratory rat, crayfish, oyster, quahog (mollusc), and fiddler crab. The purpose of Enrigh's study was to determine, using periodograms, whether the cycles of activity described by the authors of the original papers (solar, i.e. 24 h, or lunar, i.e. 24.8 h) could withstand rigorous numerical analysis.

The approach is exemplified here by a series of 28 days of observations on the perch-hopping activity of a chaffinch, a European songbird, kept under constant light conditions. The periodogram shown in Fig. 12.11a is clearly dominated by a period of 21.8 h. Figures 12.11b-d display the mean values $\bar{y}_{T,\,j}$ of the columns of the Buys-Ballot tables constructed for some of the time periods investigated: $T_k = 21.8$, 24.0 and 24.8 h. (The values $\bar{y}_{T,\,j}$ of Fig. 12.11b-d were used to calculate the amplitudes of the periodogram shown in Fig. 12.11a.) Similar figures could be drawn for each point of the periodogram, since a Buys-Ballot table was constructed for each period considered. Without the array of values in the periodogram, exclusive examination of, say, the Buys-Ballot table for $T_k = 24$ h (Fig. 12.11c) could have led to conclude to the presence of a circadian rhythm. Similarly, examination of the table for $T_k = 24.8$ h (Fig. 12.11d) could have suggested a lunar rhythm. In the present case, the periodogram allowed Enright to (1) reject periods that were intuitively interesting (e.g. $T_k = 24$ h) but whose amplitude was not significantly high, and (2) identify a somewhat unexpected 21.8-h rhythm, which seemed to be of endogenous nature.

The 17 data series re-examined by Enright (1965) had been published with the objective of demonstrating the occurrence of circadian or tidal cycles. Enright's periodogram analyses confirmed the existence of circadian cycles for *only two* of the published series: one for the rat locomotor activity, and one for the quahog shell-opening activity. *None* of the published series exhibited a tidal (lunar) cycle. This stresses the usefulness of periodogram analysis in ecology and the importance of using appropriate numerical methods when studying data series.

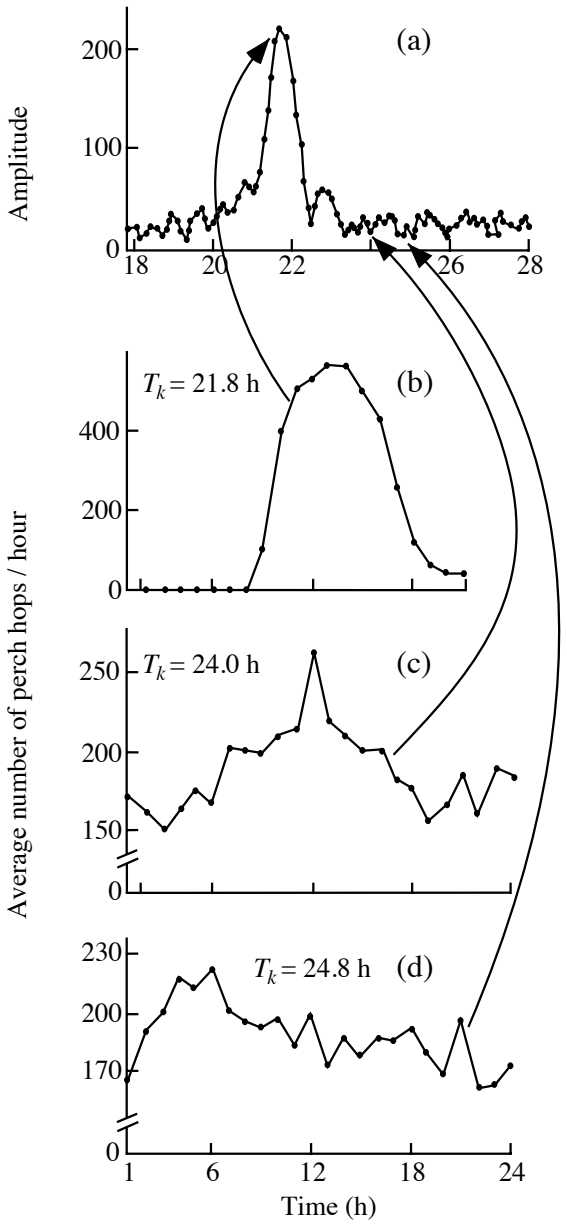


Figure 12.11 (a) Periodogram for the chaffinch perch-hopping activity series (n = 672 data points). The amplitude was calculated using Enright's formula (eq. 12.12). The three lower panels illustrate examples of values from which the amplitudes in (a) were calculated. These graphs show the means $\bar{y}_{T,\,j}$ of the columns in the Buys-Ballot tables, as functions of time, for periods T_k of (b) 21.8 h, (c) 24.0 h, and (d) 24.8 h. After Enright (1965).

Ecological application 12.4b

Nardi *et al*. (2003) used the periodogram of Whittaker & Robinson to study seasonal variations in the free-running period (i.e. circadian rhythm) in two populations of sandhopper (marine amphipods) on Italian beaches that differed in morphodynamics and human disturbance.

2 — Contingency periodogram of Legendre et al.

Another type of periodogram has been proposed by Legendre *et al*. (1981) to identify rhythms in series of *qualitative* ecological data. In this *contingency periodogram*, the Buys-Ballot table is replaced by a *contingency table* (Section 6.2). The columns of the table (Colwell, 1974) are the same as in a Buys-Ballot table, but the rows are the r states of the qualitative descriptor under study. Values in the table are frequencies f_{ij} of the states of the descriptor (rows i), observed at the various times (columns j) of period T_k. As in the periodogram of Whittaker & Robinson (above), a different table is constructed for each period T_k considered in the periodogram.

Information statistic

Information (H) as to the states of the qualitative variable of interest (S), which is accounted for by a given period T_k, is the information in common between S and the sampling axis X (most often, time). This amount of information is computed as the intersection between S and X, for period T_k:

$$H(\mathrm{S} \cap \mathrm{X}) = H(\mathrm{S}) + H(\mathrm{X}) - H(\mathrm{S}, \mathrm{X}) \qquad \textbf{(12.13)}$$

Equation 12.14 is the same as eq. 6.10, used for calculating the information shared by two descriptors (statistic B), so that $H(\mathrm{S} \cap \mathrm{X}) = \mathrm{B}$.

The *contingency periodogram* is a graph of the values $H(\mathrm{S} \cap \mathrm{X}) = \mathrm{B}$ on the ordinate, as a function of periods T_k. Periodograms, as well as spatial correlograms (Section 13.1), are often read from left (shortest periods or lags) to right (larger periods or lags). This is the case when the process that may have generated the periodic or autocorrelated structure of the data, if any, is assumed to be stronger at small lags and to generate short periods before these are combined into long periods.

Section 6.2 has shown that statistic B is related to Wilks' X^2_{W} statistic:

$$X^2_{\mathrm{W}} = 2n\mathrm{B} \quad \text{(when B in nats; eq. 6.13)}$$

or $$X^2_{\mathrm{W}} = 2n\mathrm{B} \log_e 2 = n\mathrm{B} \log_e 4 \quad \text{(when B in bits; eq. 6.14).}$$

Because X^2_{W} can be tested for significance, critical values may be drawn on the periodogram. The critical value of B is found by replacing X^2_{W} in eq. 6.13 by the critical value $\chi^2_{\alpha,\nu}$:

$$\mathrm{B}_{\mathrm{critical}} = \chi^2_{\alpha,\nu}/2n \quad \text{(for B in nats)}$$

where α is the significance level and $\nu = (r - 1)(T_k - 1)$ is the number of degrees of freedom. For the periodogram, an alternative to plotting B is to plot the χ^2_W statistic as a function of periods T_k; the critical value to be used is then $\chi^2_{\alpha,\nu}$ directly. As one proceeds from left (smaller periods) to right (larger periods) in the periodogram, T_k and ν increase; as a consequence, the critical value, $\chi^2_{\alpha,\nu}$ or $B_{critical}$, monotonically increases from left to right in this type of periodogram, as will be shown in the numerical example below.

Since multiple tests are performed in a contingency periodogram, the critical values of B must be corrected (Box 1.3). The simplest approach is the Bonferroni correction, where significance level α is replaced by $\alpha' = \alpha/$(number of simultaneous tests). In a periodogram, the number of simultaneous tests is the number of periods T_k for which the statistic (B or X^2_W) has been computed. Since the maximum number of periods that can be investigated is limited by the observational window (Section 12.0), the maximum number of simultaneous tests is $[(n/2) - 1]$ and the strongest Bonferroni correction that can be made is $\alpha' = \alpha/[(n/2) - 1]$. This is the correction recommended by Oden (1984) to assess the global significance of spatial correlograms (Section 13.1). In practice, when analysing long data series, one usually does not test the significance past some arbitrarily chosen point; if there are h statistics that have been tested for significance, the Bonferroni method would call for a corrected significance level $\alpha' = \alpha/h$.

There are two problems with the Bonferroni approach applied to periodograms and spatial correlograms. The first one is that the correction varies in intensity, depending on the number of periods (in periodograms) or lags (in spatial correlograms) for which statistics have been computed and tested. The second problem is that the interest in the results of the tests of significance decreases as the periods (or lags) get longer, especially in long data series; when a basic period has been identified, its harmonics are of lesser interest. These problems can be resolved by resorting to a *progressive Bonferroni* correction, proposed by P. Legendre in the Hewitt *et al*. (1997) paper. In this method, the first periodogram or spatial correlogram statistic is tested against the α significance level; the second statistic is tested against the Bonferroni-corrected level $\alpha' = \alpha/2$ because, at this point, two tests have been performed; and so forth until the k-th statistic, which is tested against the Bonferroni-corrected level $\alpha' = \alpha/k$. This approach also solves the problem of "where to stop computing a periodogram or spatial correlogram"; one goes on as long as significant values are likely to emerge, considering the fact that the significance level becomes progressively more stringent.

Progressive Bonferroni correction

Numerical example. Consider the following series of qualitative data ($n = 16$), for a qualitative variable with 3 states (from Legendre *et al*., 1981):

1, 1, 2, 3, 3, 2, 1, 2, 3, 2, 1, 1, 2, 3, 3, 1

To analyse period $T_k = 4$, for instance, the series is cut into segments of length 4 as follows:

1, 1, 2, 3; 3, 2, 1, 2; 3, 2, 1, 1; 2, 3, 3, 1

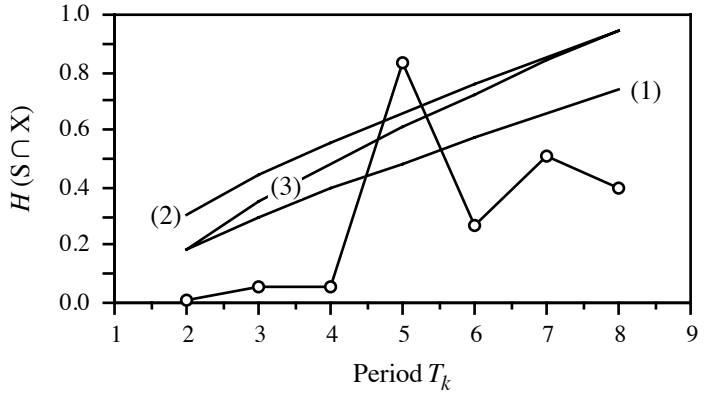


Figure 12.12 Contingency periodogram for the artificial data series (circles). The contingency statistic used here is B = H(S ∩ X). (1) Uncorrected critical values. (2) Bonferroni-corrected critical values, correcting for 7 simultaneous tests in the observational window. (3) Progressive Bonferroni correction; the first value ($T_k = 2$) is without correction, while the last ($T_k = 8$) receives the full Bonferroni correction.

and distributed in the successive rows of the table. The first four data go into columns 1 to 4 of the contingency table, each one in the row corresponding to its code; similarly, observations 5 to 8 are placed into the columns of the table, each in the row corresponding to its code; and so forth. When the operation is completed, the number of occurrences of observations are counted in each cell of the table, so that the resulting table is a contingency table containing frequencies f_{ij}. As an exercise, readers should try to reproduce the contingency tables shown below for $T_i = 4$ and $T_i = 5$. The values of X^2_W and B (in *nats*) are given for these two periods:

$T = 4$	1	2	3	4
State 1	1	1	2	2
State 2	1	2	1	1
State 3	2	1	1	1

B (in *nats*) = 0.055, $X^2_W = 1.76$

$T = 5$	1	2	3	4	5
State 1	3	3	0	0	0
State 2	1	0	3	0	1
State 3	0	0	0	3	2

B (in *nats*) = 0.835, $X^2_W = 26.72$

Repeating the calculations for $k = 2$ to 8 produces the periodogram shown in Fig. 12.12. Only $X^2_W = 26.72$ ($T = 5$) is larger than the corresponding critical value, which may be computed in various ways (as explained above), depending on the need:

• Uncorrected critical value: $\alpha = 0.05$, $\nu = (3 - 1)(5 - 1) = 8$, critical $\chi^2_{\alpha,\nu} = 15.5$. $B_{critical} = 15.5/(2 \times 16) = 0.484$.

• Bonferroni correction for 7 simultaneous tests: $\alpha' = \alpha/(n/2 - 1) = 0.05/7$, $\nu = 8$, critical $\chi^2_{\alpha',\nu} = 21.0$. $B_{critical} = 21.0/32 = 0.656$.

• Progressive Bonferroni correction. Example for the 4th test: $\alpha' = \alpha/4 = 0.05/4$, $\nu = 8$, critical $\chi^2_{\alpha',\nu} = 19.5$. $B_{critical} = 19.5/32 = 0.609$.

Thus, the only significant period in the data series is $T_k = 5$.

The contingency periodogram can be directly applied to qualitative descriptors. Quantitative or semiquantitative descriptors must be divided into states before analysis with the contingency periodogram. A method to do so is described in Legendre *et al.* (1981).

In their paper, Legendre *et al.* (1981) established the robustness of the contingency periodogram in the presence of strong random variations, which often occur in ecological data series, and its ability to identify hidden periods in series of non-quantitative ecological data. Another advantage of the contingency periodogram is its ability to analyse very short data series.

One of the applications of the contingency periodogram is the analysis of multivariate series (e.g. multi-species; Ecological application 12.4c). Such series may be transformed into a single qualitative variable describing a partition of the observations found by clustering (Chapter 8). With the contingency periodogram, it is possible to analyse the data series, now transformed into a single nonordered variable (factor) corresponding to the partition of the observations. An alternative approach would be to carry out the analysis on the multivariate distance matrix among observations using the Mantel correlogram described in Subsection 13.1.6.

Ecological application 12.4c

Phytoplankton was enumerated in a series of 175 water samples collected hourly at an anchor station in the St. Lawrence Estuary (Québec). Using the contingency periodogram, Legendre *et al.* (1981) analysed the first 80 h of that series, which corresponded to neap tides. The original data consisted of six functional taxonomic groups. The *six-dimensional quantitative data* were transformed into a *one-dimensional qualitative descriptor* by clustering the 80 observations using flexible clustering (Subsection 8.5.10). Five clusters of "hours" were obtained; each hour of the series was attributed to one of them. Each cluster thus defined a state of the new qualitative variable resulting from the classification of the hourly data.

When applied to the qualitative series, the contingency periodogram identified a significant period $T = 3$ h, which suggested rapid changes in surface waters at the sampling site. The integer multiples (harmonics) of the basic period (3 h) in the series also appeared in the contingency periodogram. Periods $T = 6$ h, 9 h, and so on, had about the same significance as the basic period, so that they did not indicate the presence of additional periods in the series.

3 — Periodogram of Schuster

Harmonic analysis

For *quantitative* serial variables, there exists another method for calculating a periodogram, which is mathematically more complex than the periodogram of Whittaker and Robinson (Subsection 12.4.1) but is also more powerful. It is sometimes called *harmonic analysis* or *periodic regression*. This method is based on the fact that

the periodic variability present in series of quantitative data can often be represented by a sum of periodic terms, involving combinations of sines and cosines (Fig. 12.13):

Fourier series

$$y(x) = a_0 + \sum_k \left[a_k \cos\left(\frac{2\pi}{T_k} x\right) + b_k \sin\left(\frac{2\pi}{T_k} x\right) \right] \quad \textbf{(12.14)}$$

Equation 12.14 is called a *Fourier series*. Constant a_0 is the mean of the series; parameters a_k and b_k determine the importance of a given period T_k in the resulting signal. Using eq. 12.14, any periodic signal can be partitioned into a sequence of superimposed oscillations (Fig. 12.13). Function $\cos[x(2\pi/T_k)]$ transforms the explanatory variable x into a cyclic variable. Periods T_k are generally chosen in such a way that the sines and cosines, which model the data series, are *harmonics* (Section 12.0) of a fundamental period T_0: $T_k = T_0/k$ (where $k = 1, 2, \ldots, n/2$). Periods T_k become shorter as k increases. Equation 12.15 may be rewritten as:

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos\left(k\frac{2\pi}{T_0} x\right) + b_k \sin\left(k\frac{2\pi}{T_0} x\right) \right]$$

Generally, T_0 is taken to be equal to the length of the series ($T_0 = n\Delta$, where Δ is the interval between data points), so that:

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos\left(k\frac{2\pi}{n\Delta} x\right) + b_k \sin\left(k\frac{2\pi}{n\Delta} x\right) \right] \quad \textbf{(12.15)}$$

The purpose of Fourier analysis is not to determine the values of coefficients a_k and b_k, but to find out which periods, among all periods T_k, best explain the variance observed in the response variable $y(x)$. After estimating the values of a_k and b_k, the *amplitude* of the periodogram for each period T_k is computed as the *fraction of the variance* of the series that is explained by the given period. This quantity, which is a sum of coefficients of partial determination, combines the estimates of coefficients a_k and b_k as follows:

$$S^2(T_k) = (a_k^2 + b_k^2)/2 \quad \textbf{(12.16)}$$

Values in the periodogram are thus calculated by fitting to the data series (by least squares) a finite number of sine and cosine functions with different periods. There are $n/2$ such functions in the harmonic case. The shortest period considered is 2Δ ($T_{k\,\max} = T_0/(n/2) = n\Delta/(n/2) = 2\Delta$). It corresponds to the lower limit (expressed in period or wavelength) of the observational window (bottom row of Table 12.1). The amplitude is computed for each period T_k independently.

Schuster periodogram

Plotting the amplitudes from eq. 12.16 as a function of periods T_k produces the *periodogram of Schuster* (1898), which is used to identify significant periods in data series. In usual calculations, frequencies T_k are harmonics of T_0, but it is also possible

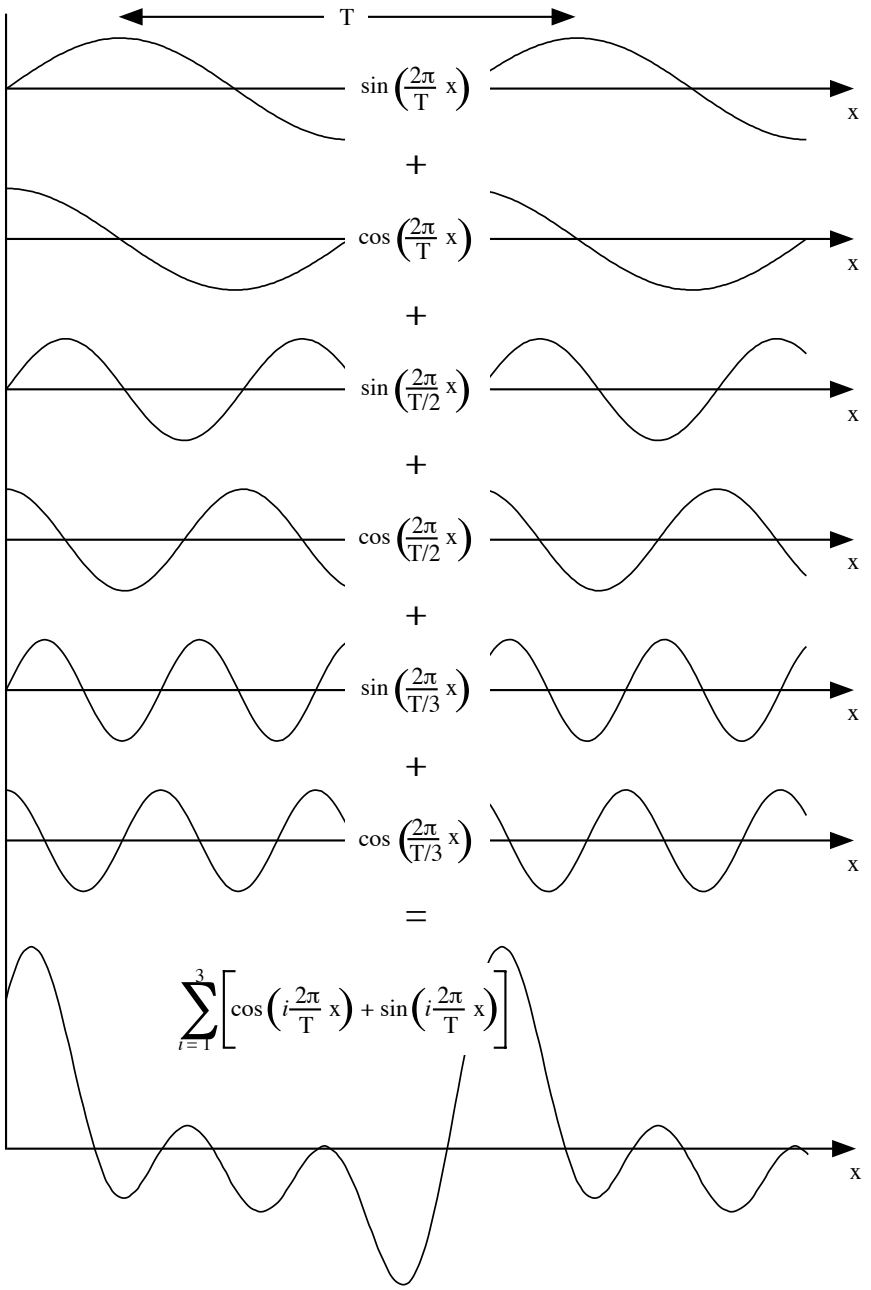


Figure 12.13 Fourier series. The periodic variation in this example (bottom graph, same as the periodic component of Fig. 12.2b) results from the sum of three sines and three cosines, which make up a harmonic sequence ($T_k = T, T/2$ and $T/3$). The mean of the series is 0 ($a_0 = 0$) and the amplitude of each sine and cosine is equal to 1 ($a_k = b_k = 1$).

to choose them to correspond to values of particular interest in the study. Contrary to the periodogram of Whittaker & Robinson, which does not refer to an underlying mathematical model, Schuster's periodogram is based on Fourier series (eqs. 12.14 and 12.15). Indeed, Kendall & Ord (1990, p. 158) have shown that any time series may be decomposed into a set of cycles based on the harmonic frequencies, even if the series does not display periodicity. Spatial eigenfunctions (Sections 14.1 to 14.3), computed along the time series, can be used for the same type of decomposition.

One advantage of Schuster's periodogram is that it can handle series showing several periods, contrary to the periodogram of Whittaker and Robinson which is limited to series with only one stable period (see above). Values in Schuster's periodogram can be tested for significance by reference to a critical value, which is calculated using a formula derived from Anderson (1971, p. 110 *et seq.*):

$$-(2/n)\log_e\left(1-\sqrt[m]{1-\alpha}\right) \qquad \textbf{(12.17)}$$

where n is the number of observations in the series, m is the largest computed harmonic period (usually, $m = n/2$), and α is the significance level.

Ecological application 12.4d

Demers & Legendre (1981) used Schuster's periodogram to analyse a 76-h series of oceanographic data. For a significance level $\alpha = 0.05$, the critical value for the periodogram was:

$$-(2/76)\log_e\left(1-\sqrt[38]{1-0.05}\right) = 0.174 = 17.4\%$$

Hence, any period explaining more than 17.4% of the variance of the series was considered to be significantly different from zero at significance level $\alpha = 0.05$.

Ecological application 12.4e

The time series of chlorophyll *a* and *Trichodesmium* filaments in tropical waters (Steven & Glombitza, 1972), discussed in Ecological applications 12.3b and 12.3c above, were subjected to harmonic analysis. Results are reported in Table 12.8. Each column of results could also be plotted as a periodogram. The period $T = 120$ days, already evidenced by autocorrelation (Fig. 12.8) and cross-correlation (Fig. 12.9), was also clearly identified by harmonic analysis.

Ecological application 12.4f

Crow birds act as a reservoir of the West Nile virus (WNV), which first appeared in North America in 1999. Ludwig *et al*. (2009) used Schuster's periodogram to investigate the population dynamics of crow birds in Québec and evaluate the impact of WNV infection on these dynamics. Their purpose was to develop a predictive algorithm that could be used as a disease surveillance tool and a measure of the impact of WNV on wildlife.

Table 12.8 Harmonic analysis of time series of chlorophyll *a* and *Trichodesmium* filaments, in tropical marine waters. The table reports the amplitudes corresponding to harmonic periods. The dominant period (T_k = 120) is in italics. After Steven & Glombitza (1972).

Harmonic	Period	Nonfiltered series		Filtered series	
k	T_k = 840 days/k	Chl *a*	*Trichodesmium*	Chl *a*	*Trichodesmium*
4	210	0.007	67	0.010	75
5	168	0.007	178	0.006	168
6	140	0.022	113	0.019	129
7	*120*	*0.039*	*318*	*0.038*	*311*
8	105	0.017	147	0.016	162
9	93	0.018	295	0.019	291
10	84	0.020	123	0.020	144

4— Periodogram of Dutilleul

Fractional periods do not correspond to an integer number of cycles in the series. These periods are usually not computed in Schuster's periodogram, although there is nothing that prevents it mathematically except the fact that the test of statistical significance of individual values (eq. 12.17) is only asymptotically valid with fractional periods. As a consequence, Schuster's periodogram is poorly adapted to the analysis of short time series, in which the periods of interest are likely to be fractional. A rule of thumb is to only analyse series that are at least 10 times as long as the longest hypothesized period.

Dutilleul (1990) proposed to modify Schuster's periodogram, in order to compute the portion of total variance associated with periods that do not correspond to integer fractions of the fundamental period T_0 (i.e. *fractional periods*). The method allows a more precise detection of the periods of interest and is especially useful with *short data series*.

Dutilleul periodogram

The statistic in *Dutilleul's modified periodogram* is the exact fraction of the total variance of the time series explained by regressing the series on the sines and cosines corresponding to one or several periodic components. In contrast, Schuster's periodogram is estimated *for a single period at a time*, i.e. each period T_k in eq. 12.14. It follows that, when applied to short series, Schuster's periodogram generally only provides an approximation of the explained fraction of the variance. In general, the number of periodic components actually present in a series is unknown *a priori*, but it may be estimated using a stepwise procedure proposed by Dutilleul (1990; see also

Dutilleul, 1998). The modified periodogram thus offers two major extensions over Schuster's: (1) it may be computed for *several periods at a time* (i.e. it is *multifrequential*) and (2) its maximization over the continuous domain of possible periods provides the maximization of the sum of squares of the corresponding trigonometric model fitted by least squares to the series. Both periodograms lead to the same estimates when computed for a single period over a long data series, or when the period corresponds to an integer fraction of T_0. In all other cases, the modified periodogram has better statistical properties (Dutilleul, 1990; see also Legendre & Dutilleul, 1992; Dutilleul & Till, 1992; Dutilleul, 1998, 2011):

- The explained fraction of the variance tends to be maximum for the true periods present in the time series, even when these are fractional, because the periodogram statistic exactly represents the sum of squares of the trigonometric model fitted by least squares to the series at the frequencies considered, whether these are integers or not (when expressed in number of cycles over the series).

- Assuming normality for the data series, the periodogram statistic is distributed like χ^2 for all periods in small or large samples, which leads to exact tests of significance. With Schuster's periodogram, this is only the case for periods corresponding to integer fractions of T_0 or, outside these periods, only for large samples.

- When the number of frequencies involved in the computation corresponds to the true number of periodic components in the series, the frequencies maximizing the periodogram statistic are unbiased estimates of the true frequencies. The stepwise procedure mentioned above allows the estimation of the number of periodic components present in the series.

In order to compare Dutilleul's periodogram to Schuster's, Legendre & Dutilleul (1992) created a test data series of 30 simulated observations containing two periodic components, which jointly accounted for 70.7% of the total variance in the series, with added noise. The true periods were $T = 12$ and 15 units. Schuster's periodogram brought out only one peak, because the two components were close to each other and Schuster's periodogram statistic was estimated for only one period at a time. When estimated for a single period, Dutilleul's modified periodogram shared this drawback. However, when estimated for the correct number of periods (i.e. two, as found by the stepwise procedure mentioned above), the modified periodogram showed maxima near the two constructed periods, i.e. at $T = 11.3$ and 14.4 units. The authors also compared the results of Dutilleul's method to those obtained with the stepwise procedure of Damsleth & Spjøtvoll (1982), which is based on Schuster's periodogram. Results from the latter (estimated periods $T = 10.3$ and 13.5) were not as good as with Dutilleul's modified periodogram. Dutilleul (1998) also showed the better performance of the modified periodogram over autocorrelograms in the context of scale analysis.

Dutilleul & Till (1992) published an application of the modified periodogram to the analysis of long dendrochronological series. Dutilleul's periodogram clearly detected the annual solar signal in cedar tree-ring series in the Atlas, a sub-tropical region

where, typically, the annual dendrochronological signal is weak. An application to a series of moderate length (river discharge) was published by Tardif *et al*. (1998).

Dutilleul (2011, his Section 6.3.2) discussed applications of Dutilleul's periodogram to several data series: maternal behaviour of the Wistar rat (a strain of albino rats) in the laboratory, observed every two hours during five days (his Fig. 6.7, c1 and c2); yearly mean sunspot numbers for the period 1749-1994 (his Fig. 6.8, d and e); monthly atmospheric CO_2 concentrations at Mona Laua, Hawaii, from 1965 through 2004 (his Fig. 6.9b); and hourly mean temperatures in air and soil in the Gault Nature Reserve (Québec) over eight days in June 2004 (his Fig. 6.10, b and f).

5— Harmonic regression

Legand (1958) proposed to use the first term of the Fourier series (eq. 12.14) to analyse ecological periodic phenomena with known *sinusoidal* periodic variability (e.g. circadian or annual). This method is called *harmonic regression*. As in the case of Fourier series (see above), the explanatory variable x (e.g. time of day) is transformed into a cyclic variable:

$$x' = \cos\left[\frac{2\pi}{T}(x + c)\right] \qquad \textbf{(12.18)}$$

for cosine functions using angles in radians, as in R. In the above expression, which is the first term of a Fourier series, T is the period suggested by hypothesis (e.g. 24 hours); x is the explanatory variable (e.g. local time); and 2π is replaced by 360° when the cosine function uses angles in degrees. Constant c fits the position of the cosine along the abscissa, so that it corresponds to the time of minimum and maximum values in the data set. The regression coefficients are estimated by the least-squares method:

$$\hat{y} = b_0 + b_1 x'$$

The harmonic regression equation can be fitted to data series by nonlinear least squares using function ***nls()*** in R (Section 10.7).

Ecological application 12.4g

Angot (1961) studied the diurnal cycle of marine phytoplankton production near New Caledonia, in the South Pacific. Values of primary production exhibited regular diurnal cyclic variations, which might reflect physiological rhythms. After logarithmic transformation of the primary production values, the author found *significant* harmonic regressions, with $T = 24$ h and $c = 3$ h; the explanatory variable x was the local time. Coefficients of regression b_0 and b_1 were used to compare different sampling sites.

Ecological application 12.4h

Taguchi (1976) used harmonic regression to study the short-term variability of marine phytoplankton production for different irradiance conditions and seasons. Data, which represented a variety of coastal conditions, were first transformed into ratios of production to chlorophyll *a*. The explanatory variable x was local time, $c = 4$ h, and T was generally 24 h. The intercept b_0 represented the mean production and b_1 was the slope of the regression line. The two coefficients decreased with irradiance and varied with seasons. The author interpreted the observed changes of regression coefficients in terms of photosynthetic dynamics.

Periodogram analysis is of interest in ecology because calculations are relatively simple and interpretation is direct. The correlogram and periodogram approaches, however, often give way to *spectral analysis* (next section). Spectral analysis is more powerful than correlogram or periodogram analyses, but it is also a more complex method for studying series. For simple problems where spectral analysis would be an unnecessary luxury, ecologists should rely on correlograms or, better, periodograms.

12.5 Periodic variability: spectral and wavelet analyses

Spectral analysis is the most advanced approach to analyse data series. The general concepts upon which it is founded are described below and illustrated by ecological applications. However, the analysis cannot be conducted without taking into account a number of theoretical and practical considerations, whose discussion exceeds the scope of the present book. Interested readers should refer, for instance, to the review papers by Platt & Denman (1975) and Fry *et al*. (1981). They may also consult the books of Bendat & Piersol (1971) and Muller & Macdonald (2002) as well as the references provided at the end of Section 12.0. Ecologists wishing to use spectral analysis are advised to consult a colleague with *practical experience* of the method.

1 — Series of a single variable

In the previous section, calculation of the Schuster periodogram involved least-squares fitting of a Fourier series to the data (eq. 12.14):

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos\left(\frac{2\pi}{T_k} x\right) + b_k \sin\left(\frac{2\pi}{T_k} x\right) \right]$$

When calculating the periodogram, the Fourier series was constructed using periods $T_k = T_0/k$. In spectral analysis, frequencies $f_k = 1/T_k$ are used instead of periods T_k. Thus, eq. 12.14 is rewritten as:

$$y(x) = a_0 + \sum_{k=1}^{n/2} \left[a_k \cos(2\pi f_k x) + b_k \sin(2\pi f_k x) \right] \qquad \textbf{(12.19)}$$

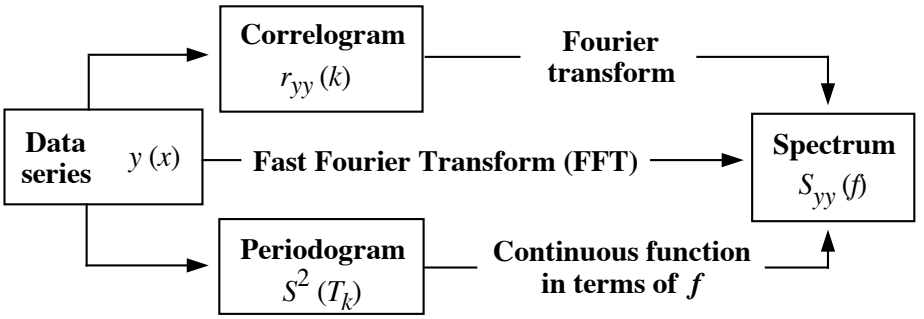


Figure 12.14 Relationships between a data series, its correlogram and periodogram, and its variance spectrum. The figure shows that the correlogram or the periodogram, on the one hand, and the spectrum, on the other hand, form a pair of Fourier transforms.

Using a formula similar to eq. 12.16, the *intensity* of the periodogram, at frequency f_k, is computed using the least-squares estimates of coefficients a_k and b_k:

$$I(f_k) = n(a_k^2 + b_k^2)/2 \qquad \textbf{(12.20)}$$

The intensity of the periodogram is defined only for *harmonic* frequencies $k/n\Delta$. It is possible, however, to turn the intensity of the periodogram into a *continuous* function over all frequencies from zero to the Nyquist frequency (see Table 12.1). This defines the *spectrum* of the series:

Spectrum

$$S_{yy}(f) = n(a_f^2 + b_f^2)/2 \qquad 0 \le f \le f_{n/2} \qquad \textbf{(12.21)}$$

The spectrum is thus a *continuous* function of frequencies, whereas the periodogram is discontinuous. Calculation and interpretation of spectra is the object of *spectral analysis*. Because of its origin in the field of electricity and telecommunications, the spectrum is sometimes called "power spectrum" or "energy spectrum". As shown below, it is also a "variance spectrum", which is the terminology used in ecology.

Power, or variance spectrum

In algebra, there exist mathematically equivalent pairs of equations that are used to go from one independent variable to another. Two mathematically equivalent equations where one is a function of x and the other a function of frequency $f = 1/x$ are called a pair of *Fourier transforms*. It can be shown that the *autocovariance* or *autocorrelation* function (eqs. 12.5-12.7) and the *spectral density* function (eq. 12.21) are a pair of Fourier transforms. In other words, the spectral density function is a Fourier transform of the autocorrelation function, and vice versa. Therefore, both the correlogram (Section 12.3) and periodogram analyses (Section 12.4), when they are generalized, lead to spectral analysis (Fig. 12.14). Classically, the spectrum is

Fast Fourier Transform

computed by Fourier transformation (also called "Fourier transform") of the autocorrelation, followed by smoothing. There is another method, called *Fast Fourier Transform* (FFT), which is faster than the classical approach (shorter computing time) and efficiently computes the pair of Fourier transforms written in discrete form. This last method offers the advantage of computational efficiency, but it involves a number of constraints, which can only be fully mastered after acquiring some *practical* experience of spectral analysis. It is sometimes confusing that, according to the context, the word "transform" is used as a verb (i.e. to transform an equation into another) or as a noun, and in the latter case it either refers to an algebraic operation (e.g. fast Fourier transform) or the result of that operation (e.g. a pairs of Fourier transforms).

Smoothing window

The spectrum computed from a correlogram or autocovariance function is an unbiased estimate of the true spectrum. However, the standard error of this spectral estimate is 100% whatever the length of the series. It follows that the computed spectrum must be *smoothed* in order to reduce its variance. Smoothing is done using a *window*, which is a function by which one multiplies the spectrum itself (spectral window), or the autocovariance estimates (lag window) prior to Fourier transformation. The two types of windows lead to the same results. The main problem of *smoothing* is that reduction of the standard error of the spectral estimates, on the ordinate, always leads to spreading of the variance on the abscissa. As a result, the spectral estimate, at any given frequency, may become contaminated by variance that is "leaking" from neighbouring frequencies. This *leakage* may result in biased smoothed spectral estimates. The various windows found in the literature (e.g. Bartlett, Daniell, de la Valle-Poussin or Parzen, Hamming, von Han, Tukey) provide different compromises between reduction of the standard error of spectral estimates and loss of resolution between adjacent frequencies. As was stressed above, the practical aspects of spectral analysis, including the choice of windows, filters (Section 12.2), and so on, often necessitate the help of an experienced colleague.

The ecological interpretation of spectra is not necessarily the same as that of correlograms or periodograms. First, the spectrum is a true *partition of the variance* of the series among frequencies. Therefore, spectral analysis is a third type of variance decomposition, in addition to the usual partitioning among experimental factors or sampling axes (ANOVA) and the partition among principal axes (Sections 4.4 and 9.1). The *units* of spectral density are [variance × frequency^{-1}], i.e. [(units of the response variable y)^2 × (units of the explanatory variable x)]. Therefore, the *variance* that corresponds to a frequency band is the *area under the curve* between the upper and lower frequencies, i.e. the integration of [variance × frequency^{-1}] over the frequency band. Spectra may be computed to identify harmonics in the data series or they may be regarded as characteristics of whole series, whether they are true sums of harmonics or not (Kendall & Ord, 1990, p. 158). These concepts should become clearer with the following ecological applications.

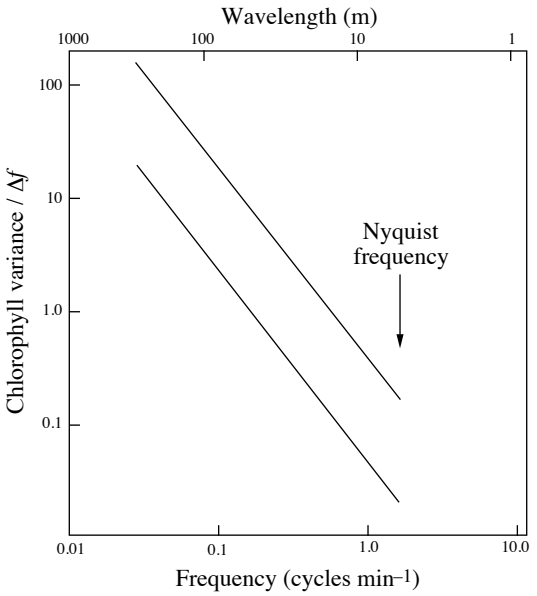


Figure 12.15 Horizontal distribution of chlorophyll *a* (*in vivo* fluorescence; arbitrary units) in surface waters of the Gulf of St. Lawrence. The two parallel lines on the variance spectrum show the envelope of calculated spectral densities. The Nyquist frequency is 1.5 cycle min^{-1}. After Platt (1972).

Ecological application 12.5a

At an anchor station in the Gulf of St. Lawrence, Platt (1972) continuously recorded *in vivo* fluorescence in surface waters as an estimate of phytoplankton chlorophyll *a*. Spectral analysis of the detrended data series (Fourier transform of autocorrelation) resulted in a spectrum characterized by a slope of –5/3, over frequencies ranging between ca. 0.01 and 1 cycle min^{-1}. The average current velocity being ca. 20 cm s^{-1} (ca. 10 m min^{-1}), the time series covered spatial scales ranging between ca. 1000 and 10 m (wavelength = speed × $frequency^{-1}$). This is illustrated in Fig. 12.15.

Interpretation of the spectrum was based on the fact that spectral analysis is a type of variance decomposition in which the total variance of the series is partitioned among the frequencies considered in the analysis (here: 0.03 cycle $min^{-1} < f <$ 1.5 cycle min^{-1}). The slope –5/3 corresponds to that of turbulent processes. This led the author to hypothesize that the local concentration of phytoplankton could be mainly controlled by turbulence. In a subsequent review paper, Platt & Denman (1975) cite various studies, based on spectral analysis, whose results confirm the hypothesis that the mesoscale spatial organization of phytoplankton is controlled by physical processes, in both marine and freshwater environments. This is in fact a modern version of the model proposed in 1953 by Kierstead & Slobodkin, which is discussed in Ecological applications 3.2d and 3.3a. Other references on spectral analysis of *in vivo* fluorescence series include Demers *et al*. (1979), Denman (1976, 1977), Denman & Platt (1975, 1976), Denman *et al*. (1977), Fashman & Pugh (1976), Legendre & Demers (1984), Lekan & Wilson (1978), Platt (1978), Platt & Denman (1975), and Powell *et al*. (1975), among others.

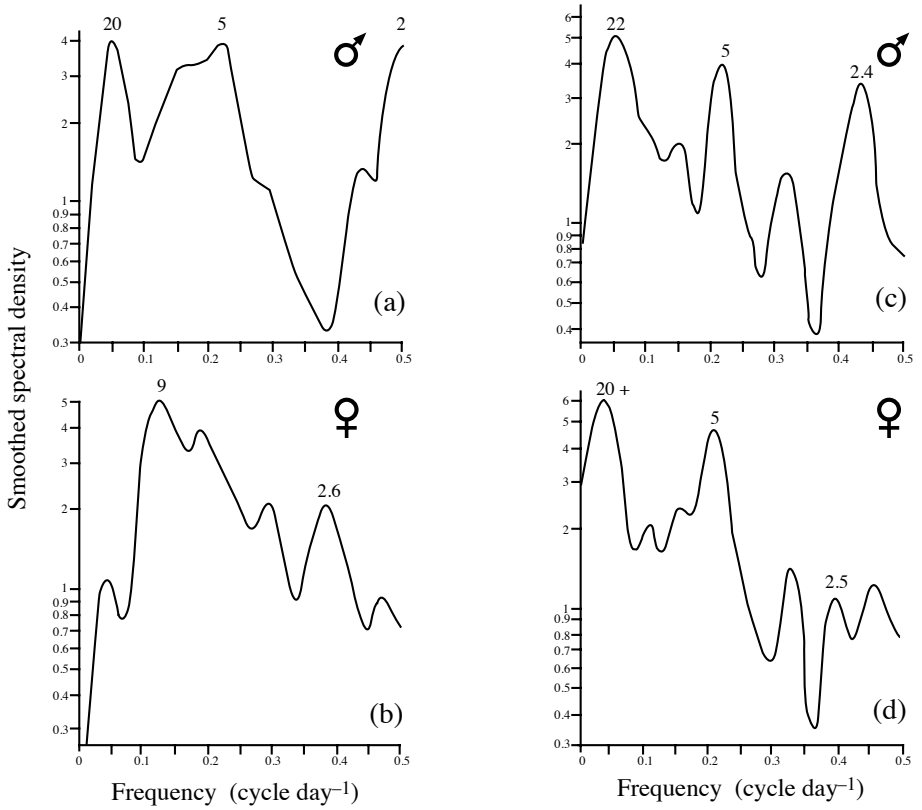


Figure 12.16 Estimates of smoothed spectra for numbers of migrating (a) male and (b) female crickets and for the locomotor activity of (c) male and (d) female crickets in the laboratory. The Nyquist frequency is 0.5 cycle day^{-1}. Periods corresponding to the main peaks are indicated above the curve, in order to facilitate interpretation; periods are the inverse of frequencies (abscissa). After Campbell & Shipp (1974).

Ecological application 12.5b

Campbell & Shipp (1974) tried to explain the migrations of an Australian cricket from observations on rhythms of locomotor activity of the males and females. One summer migration was followed during 100 days, starting in mid-February. In addition, locomotor activity rhythms of the males and females were observed in the laboratory during ca. 100 days. Figure 12.16 shows smoothed spectra for numbers of migrating crickets and locomotor activity, for both sexes.

Peaks corresponding to periods of ca. 2.5, 5, 10, and 20 days were observed in one or several spectra, which suggested a long-term biological rhythm with several harmonics. It followed from spectral analysis that the migratory waves could be explained by synchronization of the

locomotor activity cycles of individuals in the population. Migrations of the males appeared to follow a 20-day cycle, whereas those of females seemed to follow a cycle of ca. 10 days. The authors suggested that, during these periods, males attract females to their burrows and form relatively stable couples.

Ecological application 12.5c

Another ecological example, quite different from those presented above, is provided by the study of Logerwell *et al*. (1998) in the southeastern Bering Sea. There, the authors used spectral analysis to characterise the spatial aggregation patterns of thick-billed murres (*Uria lomvia*) (birds, family *Alcidae*), and their prey (e.g. juvenile fish and krill), whose biomass had been estimated by underwater acoustic surveying.

As a further example, Dutilleul (2011, his Sections 6.2.1 and 6.2.2) applied spectral analysis to time series of daily mean temperatures in air and soil in the Gault Nature Reserve (Québec) sampled over thirty days in June 2004 (his Fig. 6.3, d and h).

2 — Multidimensional series

Spectral analysis can be used not only with univariate but also with *multidimensional* series, when several ecological variables have been recorded simultaneously. This analysis is an extension of *cross-covariance* or *cross-correlation*, in the same way as the variance spectrum is a generalization of autocovariance or autocorrelation (Fig. 12.14).

Two-dimensional series

From two data series, $\mathbf{y}_j$ and $\mathbf{y}_l$, one can compute a pair of smoothed spectra S_{jj} and S_{ll} and a cross-correlation function $r_{jl}(k)$. These are used to define the *co-spectrum* (K_{jl}) and the *quadrature spectrum* (Q_{jl}):

Co-spectrum

$$K_{jl}(f) = \text{Fourier transform of } [r_{jl}(k) + r_{jl}(-k)]/2 \qquad \textbf{(12.22)}$$

Quadrature s.

$$Q_{jl}(f) = \text{Fourier transform of } [r_{jl}(k) - r_{jl}(-k)]/2 \qquad \textbf{(12.23)}$$

The *co-spectrum* (eq. 12.22) measures the distribution, as a function of frequencies, of the covariance between those components of the two series that are in phase, whereas the *quadrature spectrum* (eq. 12.23) provides corresponding information for a phase shift of 90° between the same components. For example, a sine and cosine function are in perfect quadrature. These spectra are used, below, to compute the *coherence*, *phase*, and *gain*.

The *cross-amplitude spectrum* is defined as:

Cross-amplitude spectrum

$$\sqrt{K_{jl}^2(f) + Q_{jl}^2(f)} \qquad \textbf{(12.24)}$$

The spectra for $\mathbf{y}_j$ and $\mathbf{y}_l$ are used to compute the (squared) *coherence spectrum* (C_{jl}) and the *phase spectrum* (Φ_{jl}):

Coherence spectrum

$$C_{jl}^2(f) = \frac{K_{jl}^2(f) + Q_{jl}^2(f)}{S_{jj}(f)\,S_{ll}(f)} \quad \textbf{(12.25)}$$

Phase spectrum

$$\Phi_{jl}(f) = \arctan\left(\frac{-Q_{jl}(f)}{K_{jl}(f)}\right) \quad \textbf{(12.26)}$$

The *squared coherence* (eq. 12.25) is a dimensionless measure of the correlation of the two series in the frequency domain; for frequency f, $C_{jl}^2(f) = 1$ indicates perfect correlation between two series whereas $C_{jl}^2(f) = 0$ implies the opposite. The *phase spectrum* (eq. 12.26) shows the phase shift between the two series. When the phase is a regular function of the frequency, the squared coherence is usually significantly different from zero; when the phase is very irregular, the squared coherence is generally low and not significant.

Gain spectrum

In order to assess the causal relationships between two variables, one can use the *gain spectrum* (R_{jl}^2), which is analogous to a coefficient of simple linear regression. One can determine the response of $\mathbf{y}_j$ to $\mathbf{y}_l$:

$$R_{jl}^2(f) = \frac{S_{jj}(f)\,C_{jl}^2(f)}{S_{ll}(f)} \quad \textbf{(12.27)}$$

or, alternatively, the response of $\mathbf{y}_l$ to $\mathbf{y}_j$:

$$R_{lj}^2(f) = \frac{S_{ll}(f)\,C_{jl}^2(f)}{S_{jj}(f)} \quad \textbf{(12.28)}$$

Ecological application 12.5d

In a study of the spatial variability of coastal marine phytoplankton, Platt *et al*. (1970) repeated, in 1969, the sampling programme of 1968 described in Ecological application 12.3a. This time, data were collected not only on chlorophyll *a* but also on temperature and salinity at 80 sites along a transect. Figure 12.17 shows the coherence spectra for the three pairs of series, recorded on 24 June. Strong coherence between temperature and salinity indicates that these variables well characterized the water masses encountered along the transect. Significant coherence between the series of chlorophyll *a* and those of temperature and salinity, at ca. 3 cycles (naut. mi.)$^{-1}$, were consistent with the hypothesis that the spatial distribution of phytoplankton was controlled to some extent by the physical structure of the environment.

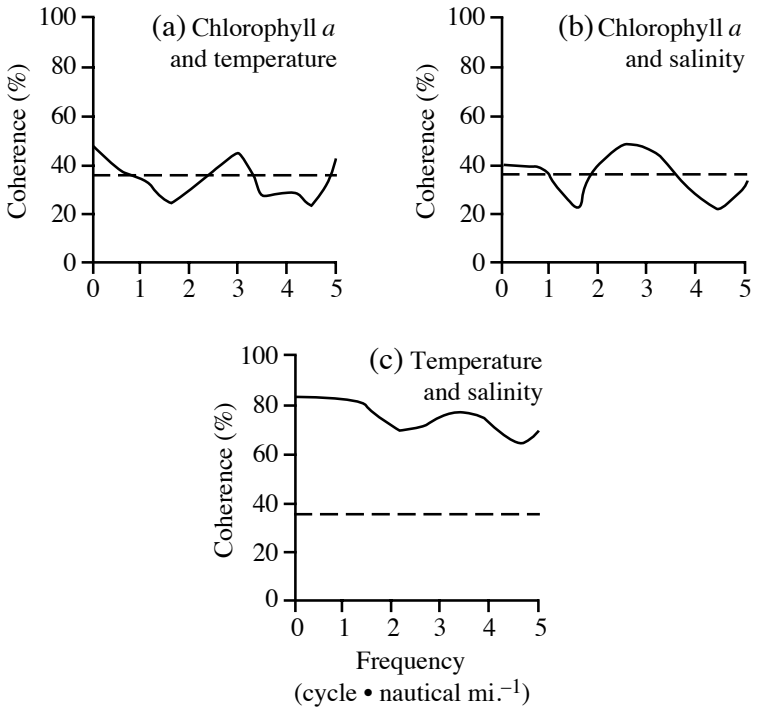


Figure 12.17 Coherence spectra between pairs of variables sampled along a transect 8 nautical miles long in St. Margaret's Bay (Nova Scotia, Canada). Dashed lines: approximate 95% confidence limits. After Platt *et al*. (1970).

Ecological application 12.5e

In order to identify the factors controlling the spatial heterogeneity of marine phytoplankton (patchiness), Denman & Platt (1975) analysed values of chlorophyll *a* and temperature, recorded continuously along a transect in the St. Lawrence Estuary. Two pumping systems were towed, at depths of 5 and 9 m, over a distance of 16.6 km (10 nautical miles). The sampling interval was 1 s, which corresponds to 3.2 m given the speed of the ship. After detrending, computations were carried out using the fast Fourier transform. Four coherence and phase spectra were calculated, as shown in Fig. 12.18.

For a given depth (Fig. 12.18a: 5 m; b: 9 m), the coherence between temperature and chlorophyll *a* was high at low frequencies and the phase was relatively constant. At higher frequencies, the coherence decreased rapidly and the phase varied randomly. The lower panels of Fig. 12.18 indicate the absence of covariation between series from different depths. The authors concluded that physical processes played a major role in the creation and control of phytoplankton heterogeneity at intermediate scales (i.e. from 50 m to several kilometres). Weak coherence between series from the two depths, which were separated by a vertical distance of only 4 m, suggested the presence of a strong vertical gradient in the physical structure. Such

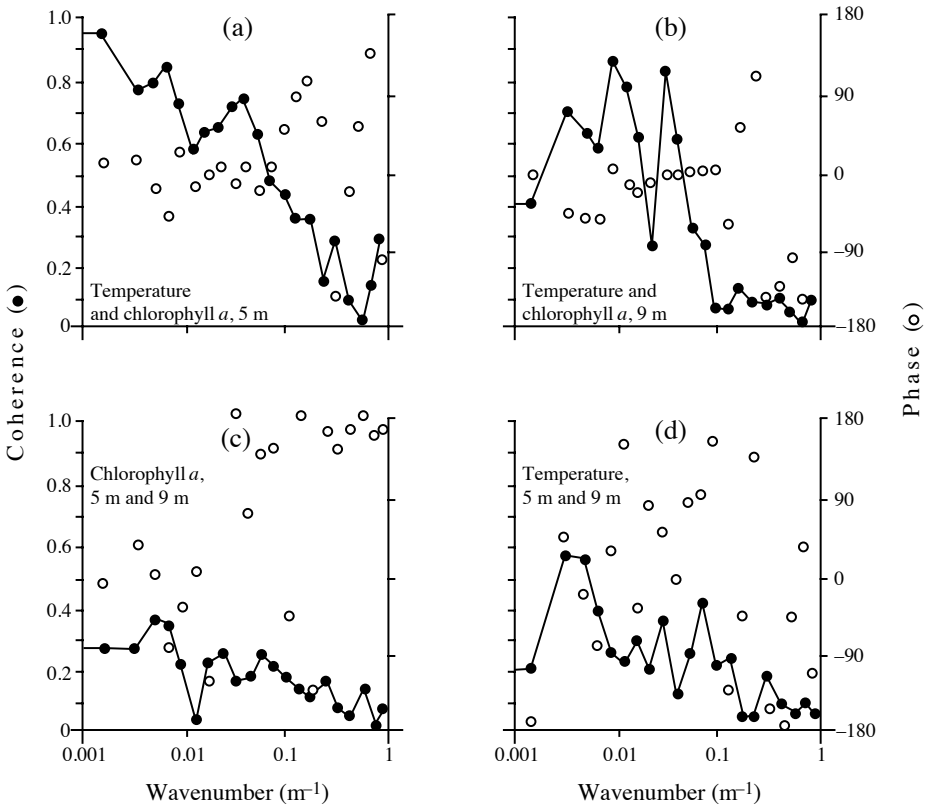


Figure 12.18 Values of coherence (solid lines) and phase (open circles), for pairs of spatial series continuously recorded in the St. Lawrence Estuary. Abscissa: *wavenumber* (= 2π/wavelength = 2π frequency/speed). Adapted from Denman & Platt (1975).

gradients are known to favour the propagation of internal waves (analogous to the propagation of waves at the air-water discontinuity). The authors proposed that the strong coherence between temperature and chlorophyll *a*, at each of the sampled depths, could reflect the presence of internal waves.

Ecological application 12.5f

In the study on the spatial distributions of thick-billed murres (*Uria lomvia*) and their prey (acoustic data) in the southeastern Bering Sea, described in Ecological application 12.5c, Logerwell *et al*. (1998) also used phase and coherence spectra. With these spectra, the authors compared the distribution patterns of birds and prey over a wide range of spatial scales.

Multivariate spectral analysis
Frequency regression

In the last paragraphs, the approach to multidimensional situations was to consider two series at a time. Brillinger (1981) provides the mathematical bases for processing multidimensional series using methods that are fully multivariate. When a stochastic series is a time-invariant function of several other series, the method recommended is *frequency regression*. It is analogous to multiple linear regression (Subsection 10.3.3), computed in the frequency domain. More generally, the method to study relationships among several series is that of *principal components in the frequency domain* (see Ecological application 12.5g). In that case, a spectrum is computed for each of the principal components, which are linear combinations of the serial variables (Section 9.1). The method has been adapted by Laurec (1979), who explained how to use it in ecology.

Another approach to the analysis of multivariate data series is the *Mantel correlogram* (Subsection 13.1.6). Because this type of correlogram is based upon a similarity or distance matrix among observations (Chapter 7), it is suitable to analyse multivariate data. It can also be used to analyse univariate or multivariate series of *semiquantitative*, *qualitative*, or *binary* data, like the species presence-absence data often collected by ecologists. Yet another approach is spatial eigenfunction analysis (Chapter 14). The study of a sediment core representing 10000 years of sedimentation (101 levels, 139 diatom species) by Legendre & Birks (2012), reported near the end of Subsection 14.1.3, is an example of analysis of a multivariate ecological series.

Ecological application 12.5g

Arfi *et al*. (1982, pp. 359-363) reported results from a study on the impact of the main sewage effluent of the city of Marseilles on coastal waters in the Western Mediterranean. During the study, 31 physical, chemical, and biological variables were observed simultaneously, at an anchor station 1 km offshore, every 25 min during 24 h ($n = 58$). Spectra for individual series (detrended) all showed a strong peak at T = ca. 6 h. Comparing the 31 data series two at a time would not have made sense because this would have required $(31 \times 30)/2 = 465$ comparisons. Thus, the 31-dimensional data series was subjected to principal component analysis in the frequency domain. Figure 12.19 shows the 31 variables, plotted in the plane of the first two principal components (as in Fig. 9.5), for $T = 6$ h. The long arrows pointing towards the upper left-hand part of the graph corresponded to variables that were indicative of the effluent (e.g. dissolved nutrients, bacterial concentrations) whereas the long arrows pointing towards the lower right-hand part of the ordination plane corresponded to variables that indicated unperturbed marine waters (e.g. salinity, dissolved O_2, phytoplankton concentrations). The positions of the two groups of variables in the plane show that their variations were out of phase by ca. 180°, for period $T = 6$ h. This was interpreted as a periodic increase in the effluent every 6 h. This periodicity corresponded to the general activity rhythm of the adjacent human population (wake-up, lunch, end of work day, and bedtime).

3 — Maximum entropy spectral analysis

As explained in Subsection 12.5.1, estimating spectra requires the use of spectral or lag *windows*. Each type of window provides a compromise between reduction of the standard error of the spectral estimates and loss of resolution between adjacent

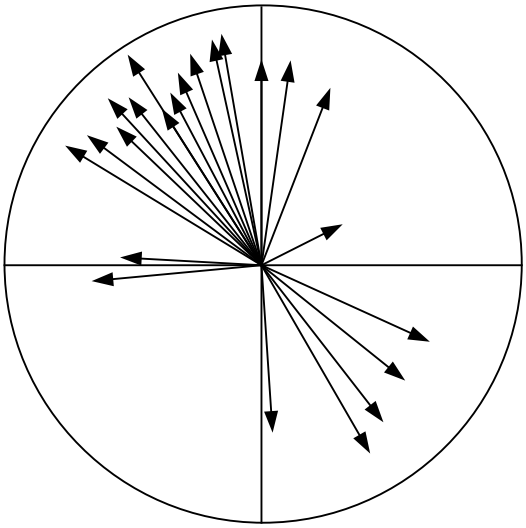


Figure 12.19 Principal component analysis in the frequency domain of 31 simultaneous series of physical, chemical, and biological variables, obtained at an anchor station in the Western Mediterranean. Plot of the 31 variables (arrows), in the plane of the first two principal components, for period $T = 6$ h. Adapted from Arfi *et al*. (1982).

frequencies. As an alternative to windows, Burg (1967) proposed to improve the spectral resolution by *extrapolating* the autocorrelation function beyond the maximum lag ($k_{\max}$), whose value is limited by the length of the series (Subsection 12.3.1). For each extrapolated lag ($k_{\max} + k$), he suggested to calculate an autocorrelation value $r_{yy}(k_{\max} + k)$ that *maximizes the entropy* (Chapter 6) of the probability distribution of the autocorrelation function. Burg's (1967) method will not be further discussed here, because a different algorithm (Bos, 1971; see below) is now used for computing this *maximum entropy spectral analysis* (MESA). Estimation of the spectrum, in MESA, does not require spectral or lag windows. An additional advantage, especially for ecologists, is that it allows the computation of spectra for very short series.

AR model

Data series may be mathematically described as stochastic linear processes. A corresponding mathematical model is the *autoregressive model* (also called *AR model* or *all-pole model*), where each observation in the series $\tilde{y}_t$ (centred on the mean $\bar{y}$ of the series: $\tilde{y}_t = y_t - \bar{y}$) is represented as a function of the q preceding observations:

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \ldots + \phi_q \tilde{y}_{t-q} + a_t \qquad \textbf{(12.29)}$$

q specifies how many steps back one takes into account to forecast value $\tilde{y}_t$. This is called the *order* of the process. The *autoregression coefficients* ϕ are estimated using

the observations of the data series itself. Residual values a_t must be independent of one another; the series of residual values is called *white noise*. Their overall variance is noted s_a^2. This type of model will be further discussed in Section 12.7.

White noise

Concerning maximum entropy spectral analysis, Bos (1971) has shown that the maximum entropy method proposed by Burg (1967) is equivalent to a least-squares fitting of an AR model to the data series. Using the autoregression coefficients ϕ, it is possible to compute the same spectral densities as those resulting from the entropy calculation of Burg (1967). Thus, the spectrum is estimated directly from the autoregression coefficients ϕ of the AR model, which are themselves estimated from the values $\tilde{y}_t$ of the data series. The spectral density for each frequency f is:

$$S(f) = \frac{2s_a^2\Delta}{\left|1 - \sum_{j=1}^{q} \phi_j \exp(-2i\pi fj)\right|^2} \qquad \textbf{(12.30)}$$

where $i = \sqrt{-1}$. Generally, the sampling interval is $\Delta = 1$ time or space unit.

Maximum entropy spectral analysis is not entirely free of problems. Some of these are still the subject of specialized papers. A first practical problem is choosing the *order* q of the AR model for an empirical data series. Various criteria for determining q have been reviewed by Berryman (1978) and Arfi & Dumas (1990). Another problem concerns the estimation of the coefficients of the AR model (see, for instance, Ulrych & Clayton, 1976). A third problem, also discussed by Ulrych & Clayton (1976), is that other processes may fit the data series better than the AR model; for example, an autoregressive-moving average model (ARMA; Section 12.7). Fitting such models may, however, raise other practical problems. The criteria for deciding to use models other than AR are partly intuitive (Section 12.7).

Ulrych & Bishop (1975) briefly reviewed the theoretical bases underlying the algorithms of Burg (1967) and Bos (1971). Barrodale & Erikson (1980) propose another algorithm for estimating the coefficients ϕ of the AR model, based on least squares, which provides a more precise estimation of the spectrum frequencies. The same authors criticize, on an empirical basis, the method of Akaike, and they propose a different approach.

Maximum entropy spectral analysis can handle short series as well as series with data exhibiting measurement errors (Ables, 1974). It may also be used to analyse series with missing data (Ulrych & Clayton, 1976). Arfi & Dumas (1990) compared MESA to the classical Fourier approach, using simulated and real oceanographic data series. For long series ($n = 450$), the two approaches have the same efficiency when noise is low, but MESA is more efficient when noise is high. For short ($n = 49$ to 56) and very short ($n = 30$) series, MESA is systematically more efficient. For long data series with low noise, it may often be simpler to compute the spectrum in the traditional way (Berryman, 1978). However, for many ecological data series, MESA would be the

method of choice. The maximum entropy approach can be generalized to handle multivariate series, since coherence and phase spectra can be computed (Ulrych & Jensen, 1974).

Spectral analysis and, thus, Objective 3 of the analysis of data series (Table 12.2), are presently restricted to *quantitative data*. The only exception is the computation of spectra for long (i.e. $n > 500$ to 1000) series of *binary variables*, using the method of Kedem (1980). Since MESA is not very demanding as to the precision of the data, it could probably be used as well for analysing series of *semiquantitative data* coded using several states.

Ecological application 12.5h

Colebrook & Taylor (1984) analysed the temporal variations of phytoplankton and zooplankton series recorded monthly in the North Atlantic Ocean and in the North Sea during 33 consecutive years (1948 to 1980). Similar series were also available for some environmental variables (e.g. surface water temperature). The series were analysed using MESA. In addition, coherence spectra were computed between series of some physical variables and the series representing the first principal component calculated for the plankton data. For the plankton series, one spectrum was computed for each species in each of 12 regions, after which the spectra were averaged over the species in each region. The resulting 12 species-averaged spectra exhibited a number of characteristic periods, of which some could be related to periods in the physical environment using coherence spectra. For example, a 3 to 4-year periodicity in plankton abundances was associated to heat exchange phenomena at the sea surface. Other periods in the spectra of the physical and biological variables could not easily be explained. Actually, 33-year series are relatively short compared with the long-term meteorological or oceanographic variations, so that some of the identified periods may turn out not to be true cycles.

Ecological application 12.5i

Kim *et al*. (2003) measured the oxygen consumption rates of sublittoral-dwelling Washington clams (*Saxidomus purpuratus*) collected in southern South Korea. Using MESA, they evidenced two endogenous rhythms in clam respiration kept under constant conditions, i.e. during 7-9 days after collection. They found a rhythm that corresponded to the tides in their original environment, followed by a shift to a circadian rhythm.

4 — Wavelet analysis

Subsection 12.5.1 introduced the notion of pairs of mathematically equivalent equations that are called pairs of transforms, and applied it to Fourier transforms. It was then shown that the autocovariance or autocorrelation function (eqs. 12.5-12.7) and the spectral density function (eq. 12.21) are a pair of Fourier transforms. Another type of transform, called wavelet transform, provides a somewhat different approach to the analysis of data series, including ecological series. Although the wavelet transform can be regarded as a generalisation of the Fourier transform, the former may be better adapted to ecological data series than the latter (Cazelles *et al*., 2008). This is because Fourier analysis decomposes the signal into waveforms that have constant

amplitude along the time axis (i.e. the sines and cosines in Fig. 12.13), whereas wavelet analysis uses waveforms (wavelets) that are narrow when the features of the signal are high-frequency and occur over a short period along the time axis, and wide when these features are low-frequency and occur over a long period. In practice, the wavelet transform decomposes the signal over functions (called wavelets) that are narrow in the portions of the data series presenting high-frequency features, and wide where structures in the data series are of low frequency.

Section 12.3 explained that a basic assumption of the correlation-based techniques used in series analysis is stationarity, i.e. the statistical parameters of a stationary time series are constant along the time axis. However, many ecological processes violate the stationarity assumption, including population dynamics (e.g. Cazelles & Hales, 2006). As explained by various authors including Cazelles *et al*. (2008), wavelet analysis overcomes the problems of non-stationarity in time series by performing local time-scale decomposition of the signal, i.e. it estimates different spectral characteristics along the time axis. As in the case of Fourier analysis for multidimensional series (Subsection 12.5.2), it is possible to investigate relationships between two signals using wavelet cross-spectrum and coherence.

In practice, wavelet analysis is only useful to analyse univariate, regular data without gaps. For one-dimensional time series or spatial transects, the data set must be fairly large, the time interval between neighbouring observations (i.e. the lag) must be small, and the series must be long compared to the structures to be extracted. In the context of spatial analysis (Chapter 13), wavelets can be used for the analysis of two-dimensional data on a grid, e.g. remotely sensed data, or forest plots that have been entirely studied; see note in Subsection 6.5.3 about the CTFS permanent forest plots and Ecological application 14.1b where data from one of those plots are analysed.

Basic principles of wavelet analysis, and applications to both artificial data series and real ecological time series, are found in Cazelles *et al*. (2008). In that paper, the authors analyse real ecological time series describing fluctuations in populations of red grouse in Scotland over 100 years, and the association between sunspot numbers and populations of lynx and porcupine over almost 200 years.

Fortin & Dale (2005, their Section 2.6.6) provide a short introduction to wavelet analysis. Readers may refer to Dale & Mah (1998), Percival & Walden (2000), and Keitt & Urban (2005) for more in-depth introductions to this type of analysis. Analysis of ecological time series with the wavelet approach offers a new perspective for the treatment of univariate data series that do not meet the stationarity assumption. For that reason, the number of publications reporting analyses of this kind is rapidly growing.

12.6 Detection of discontinuities in multivariate series

Ecological succession

Detection of discontinuities in *multivariate data series* is a problem familiar to ecologists (Objective 4 of Section 12.1 and Table 12.2). For example, studies on changes in species assemblages over time often refer to the concept of *ecological succession*. According to Margalef (1968), the theory of species succession within ecosystems plays the same role in ecology as evolutionary theory does in general biology.

The simplest way to approach the identification of discontinuities in multivariate series is by *visual inspection* of the curves depicting changes with time (or along a spatial direction) in the abundance of the various taxa or/and in the values of the environmental variables. In most instances, however, simple visual examination of a set of graphs does not allow one to unambiguously identify discontinuities in multivariate series. Numerical techniques must be used.

Methods of series analysis described in Sections 12.3 to 12.5 are not appropriate for detecting discontinuities in multivariate series, because the presence of discontinuities is not the same as periodicity in the data. Four types of methods are summarized here.

Instead of dividing multivariate series into subsets, Orlóci (1981) proposed a multivariate method for identifying successional trends and separating them into monotonic and cyclic components. That method may be viewed as complementary to those described below.

1 — Ordinations in reduced space

Several authors have used *ordinations in reduced space* (Chapter 9) to represent multispecies time series in low-dimensional space. To help identify the discontinuities, successive observations of the time series are connected with lines, as in Figs. 9.20 and 12.24. When several observations corresponding to a bloc of time are found in a small part of the reduced space, they may be thought of as a "step" in the succession. Large jumps in the two-dimensional ordination space are interpreted as discontinuities. This approach has been used, for example, by Williams *et al*. (1969; vegetation, principal coordinates), Levings (1975; benthos, principal coordinates), Legendre *et al*. (1984a; benthos, principal components), Dessier & Laurec (1978; zooplankton, principal components and correspondence analysis), and Sprules (1980; nonmetric multidimensional scaling; zooplankton; Ecological application 9.4a). In studies of annual succession in temperate or polar regions, using ordination in reduced space, one expects the observations to form some kind of a circle in the plane of the first two axes, since successive observations are likely to be close to each other in the multidimensional space, due to climate forcing (temporal correlation, Section 1.1), and the community structure is expected to come back to its original structure after one year; the rationale for this null model of succession is developed in Legendre *et al*.

(1985, Appendix D). Departures from a regular circular pattern are thus interpreted as evidence for the existence of subsets in the data series. In simple situations, such subsets are indeed observed in the plane of the first two ordination axes (e.g. Figs. 9.20 and 12.24). When used *alone*, however, this approach has two major drawbacks.

- Plotting a multivariate data series in two or three dimensions only is not the best way of using the multivariate information. In most studies, the first two principal axes used to represent the data series account together for only 10 to 50% of the multivariate information. In such cases, distances from the main clusters of observations to isolated objects (which are in some particular way different from the major groups) are likely to be expressed by some minor principal axes that are orthogonal (i.e. perpendicular in multidimensional space) to the main projection plane. As a consequence, these objects may be projected, in the reduced-spaced ordination, within a group from which they are actually quite different. Moreover, it has been observed that the "circle" of observations (see previous paragraph) may be deformed in a spoon shape so that groups that are distinct in a third or higher dimension may be packed together in some part of the two-dimensional ordination plane. These problems are common to all ordinations when used alone for the purpose of group recognition. They are not as severe for ordinations obtained by nonmetric multidimensional scaling, however, because that method is, by definition, more efficient than others at flattening multidimensional phenomena into a user-determined small number of dimensions (Section 9.4). The best way to eliminate this first drawback is to associate ordination to clustering results, as explained in Section 10.1. This was the approach of Allen *et al.* (1977) in a study of the phytoplankton succession in Lake Wingra. See also Fig. 12.24.

- The second drawback is the lack of a criterion for assigning observations to groups in an ordination diagram. As a consequence, groups delineated on published ordination diagrams often look rather arbitrary.

2 — Segmenting data series

Hawkins & Merriam (1973, 1974) proposed a method for segmenting a multivariate data series into homogeneous units, by *minimizing the variability* within segments in the same way as in *K*-means partitioning (Section 8.8). Their work followed from the introduction of a contiguity constraint in the grouping of data by Fisher (1958), who called it *restriction* in space or time. The method of Hawkins & Merriam has been advocated by Ibanez (1984) for studying successional steps.

Contiguity constraint

The method has three interesting properties. (a) The multidimensional series is partitioned into homogeneous groups using an *objective clustering criterion*. (b) The partitioning is done with a *constraint of contiguity* along the data series. Within the context of series analysis, contiguity means that only observations that are neighbours along the series may be grouped together. The notion of contiguity has been used by several authors to resolve specific clustering problems: temporal contiguity (Subsection 12.6.5, below) or spatial contiguity (Subsection 13.3.2). (c) The observations do not have to be equispaced.

A first problem with Hawkins & Merriam's method is that users must determine the number of segments that the method is requested to identify. To do so, the increase in explained variation relative to the increase in the number of segments is used as a guide. Any one of the stopping rules used with *K*-means partitioning could also be used here (end of Section 8.8). A solution to this problem is described in Subsection 12.6.4. For community composition data, a second problem is that strings of zeros in multispecies series are likely to result in segments that are determined by the simultaneous absence of species. That problem can be resolved by transforming the species data using one of the transformations described in Section 7.7.

3 — Webster's method

Webster (1973) proposed a rather simple method to detect discontinuities in data series. He was actually working with spatial transects, but his method is equally applicable to time series. Draw the sampling axis as a line and imagine a window that travels along that line, stopping at the mid-points between adjacent observations (if these are equispaced). Divide the window in two equal parts (Fig. 12.20a). There are observations in the left-hand and right-hand halves of the window. Calculate the difference (see below) between the points located in the left-hand and right-hand halves and plot these differences in a graph, as the window is moved from one end of the series to the other (Fig. 12.20c, d). The principle of the method is that the difference should be large at points where the left-hand and right-hand halves of the window contain values that are appreciably different, i.e. where discontinuities occur in the series. The following statistics may be used in the computations:

Window

- For univariate data, calculate the absolute value of the difference between the means of the values in the left-hand and right-hand halves of the window: $\text{Statistic} = \left|\bar{x}_1 - \bar{x}_2\right|$.

- For univariate data again, one may choose to compute the absolute value of a *t*–statistic comparing the two halves of the window: $\text{Statistic} = \left|\bar{x}_1 - \bar{x}_2\right| / s_{\bar{x}_1 - \bar{x}_2}$. If one assumes second-order stationarity of the series and uses the standard deviation of the whole series as the best estimate of the standard deviations in the two halves, this statistic is linearly related to the previous one. Alternatively, one could use the regular *t*-statistic formula for *t*-tests, estimating the variance in each window from the few values that it contains; this is not recommended as it produces values of the *t*-statistic that cannot be compared, and unstable estimates when windows are narrow, which is often the case with this method.

- For multivariate series, compare the two halves of the window using either the Mahalanobis generalized distance (D_5 or D_5^2, eq. 7.39), which is the multivariate equivalent of a *t*-statistic, or the coefficient of racial likeness (D_{12}, eq. 7.52).

The width of the window is an empirical decision made by the investigator. It is recommended to try different window widths and compare the results. The window width is limited, of course, by the spacing of observations, considering the

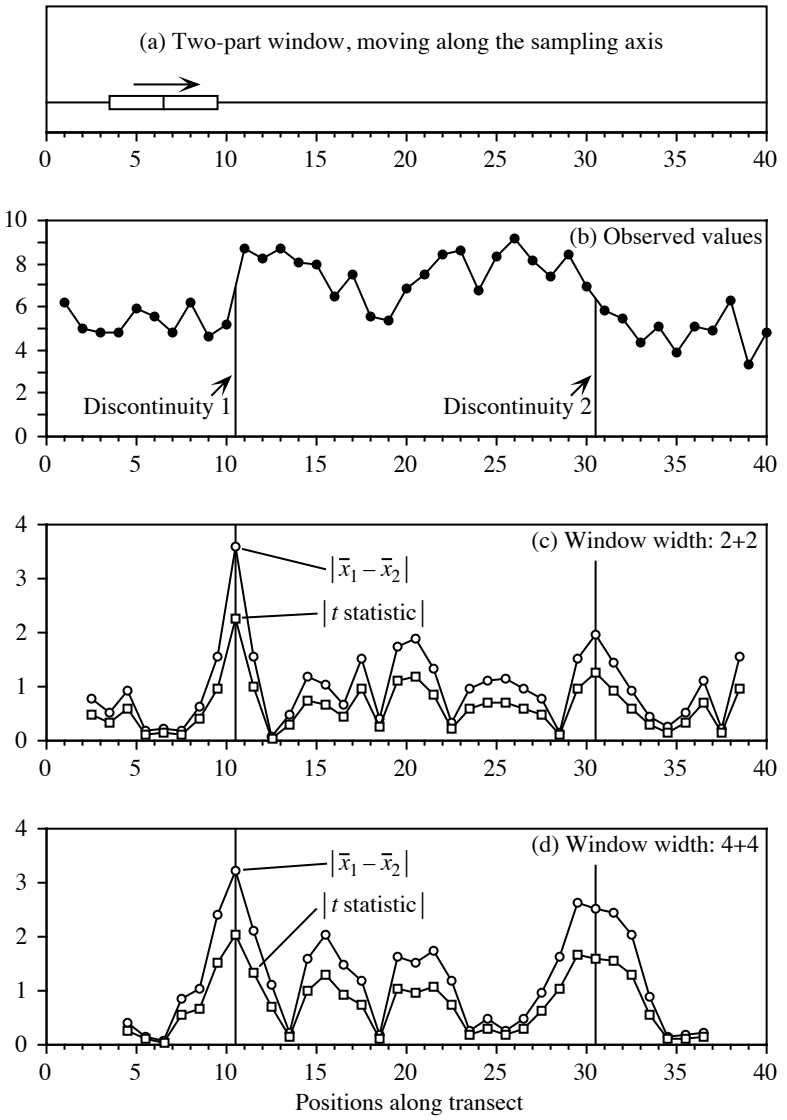


Figure 12.20 Webster's method for detecting discontinuities in data series. (a) Principle of the method. (b) Numerical example (see text). Results using a window that was (c) 4 observations wide, or (d) 8 observations wide.

approximate interval between the expected discontinuities. Webster's method works best with equispaced observations, but some departure from equal spacing, or missing data points, are allowed, because of the empirical nature of the method.

Numerical example. A series of 40 observations was generated using a normal pseudo-random number generator $N(5,1)$. The values of observations 11 to 30 were increased by adding 3 to the generated values in order to artificially create discontinuities between observations 10 and 11, on the one hand, and observations 30 and 31, on the other. It so happened that the first of these discontinuities was sharp whereas the second was rather smooth (Fig. 12.20b).

Webster's method for univariate data series was used with two window widths. The first window had a width of 4 observations, i.e. 2 observations in each half; the second window had a width of 8 observations, i.e. 4 in each half. Both the absolute values of the differences between means and the absolute values of the *t*-statistics were computed. The overall standard deviation of the series was used as the denominator of *t*, so that this statistic was a linear transformation of the difference-between-means statistic. Results (Fig. 12.20c, d) are reported at the positions occupied by the centre of the window.

The sharp discontinuity between observations 10 and 11 was clearly identified by the two statistics and window widths. This was not the case for the second discontinuity, between observations 30 and 31. The narrow window (Fig. 12.20c) estimated its position correctly, but did not allow one to distinguish it from other fluctuations in the series, found between observations 20 and 21 for instance (remember, observations are randomly-generated numbers; so there is no structure in this part of the series). The wider window (Fig. 12.20d) brought out the second discontinuity more clearly (higher values of the statistics), but its exact position was no longer estimated precisely.

D_5^2 to the centroid

Window

Ibanez (1981) proposed a related method to detect discontinuities in multivariate records (e.g. simultaneous records of temperature, salinity, *in vivo* fluorescence, etc. in aquatic environments). He called the method *D_5^2 to the centroid*. For every sampling site, the method computes a generalized distance D_5^2 (eq. 7.39) between the new multivariate observation and the centroid (i.e. multidimensional mean) of the m previously recorded observations, m defining the width of a window. Using simulated and real multivariate data series, Ibanez showed that changes in D_5^2 to the centroid, drawn on a graph like Figs. 12.20c or d, allowed one to detect discontinuities. For multi-species data, however, the method of Ibanez suffers from the same drawback as the segmentation method of Hawkins & Merriam: since the simultaneous absence of species is taken as an indication of similarity, it could prevent changes occurring in the frequencies of other species from producing high, detectable distances. That problem can be resolved by transforming the community composition data, prior to the analysis, using one of the transformations described in Section 7.7.

McCoy *et al*. (1986) proposed a segmentation method somewhat similar to that of Webster, for species occurrence data along a transect. A matrix of Raup & Crick similarities is first computed among sites (S_{27}, eq. 7.31) from the species presence-absence data. A "+" sign is attached to a similarity found to be significant in the upper tail (i.e. when a_{hi} is significantly larger than expected under the random sprinkling hypothesis) and a "–" sign to a similarity that is significant in the lower tail (i.e. when a_{hi} is significantly smaller than expected under that null hypothesis). The number of significant pluses and minuses is analysed graphically, using a rather complex empirical method, to identify the most informative boundaries in the series.

4 — Time-constrained clustering by MRT

Multivariate regression tree analysis (MRT, Section 8.11) can be used as a form of time-constrained clustering. The solution consists in analysing a multivariate response matrix **Y** using a quantitative or rank-ordered variable **x** representing the sampling sequence through time. **Y** may contain community composition data transformed in some appropriate way (Section 7.7). For a weekly time series over a year, for example, the constraining variable **x** may be a vector containing the sampling dates, counted from January 1st, or the numbers 1 to 52; the results will be identical since MRT segments **Y** at cutting points along the explanatory, or constraining, variable **x**. The observations do not have to be equispaced.

MRT is a least-squares algorithm. In the present application, it segments **Y** in such a way that the sum of the within-group multivariate sums of squares is minimum, with the constraint that the sampling dates within each group be adjacent along the sampling sequence. As a consequence, the solution obeys the Hawkins & Merriam criterion described in Subsection 12.6.2. As a bonus, the cross-validation procedure available in MRT helps determine the 'best' number of groups for the data under study; this solves the first problem of the Hawkins & Merriam method mentioned in Subsection 12.6.2. MRT can be used to segment spatial series, e.g. transect data as shown in the following ecological application, as well as time series.

Ecological application 12.6a

Borcard *et al*. (2011, their Section 4.11) used MRT to segment fish assemblage data collected at 29 sites along the Doubs River in eastern France (29 sites, 27 species) by space-constrained clustering along the course of the river. These data were also used in Ecological application 11.1a. In the present application, the data were chord-transformed (eq. 7.67) before MRT analysis. Cross-validation in MRT suggested 5 groups as the best solution; that solution had the smallest CVRE value (eq. 8.23). The five groups are represented on a map of the river in Fig. 12.21. The calculations were done with the R code provided by Borcard *et al*. (2011).

5 — Chronological clustering

Temporal contiguity

Combining some of the best aspects of the methods described above, Gordon & Birks (1972, 1974) and Gordon (1973) introduced a constraint of temporal contiguity in a variety of clustering algorithms to study pollen stratigraphy. Analysing bird surveys repeated at different times during the breeding season, North (1977) also used a constraint of temporal contiguity to cluster bird presence locations on a geographic map and delineate territories. Applications of time-constrained clustering to palaeoecological data (where a spatial arrangement of the observations corresponds to a time sequence) can be found in Bell & Legendre (1987), Hann *et al*. (1994) and Song *et al*. (1996). Algorithmic aspects of constrained clustering are discussed in Subsection 13.3.2.

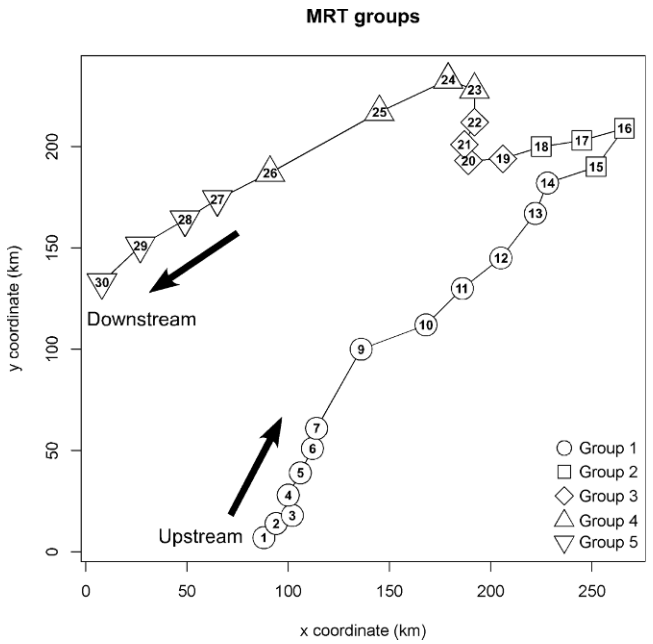


Figure 12.21 Clustering with spatial contiguity constraint of the Doubs River fish assemblage data by multivariate regression tree analysis (MRT). The groups of sites are represented on a map along the course of the river; the arrows indicate flow direction. Sites are numbered 1 to 30; site 8 was removed from the analysis for the reason explained in Ecological application 11.1a.

Succession model

Using the same concept, Legendre *et al*. (1985) developed the method of *chronological clustering*, based on hierarchical clustering (Chapter 8). The algorithm was designed to identify discontinuities in multi-species time series. It has also been successfully used to analyse spatial transects (e.g. Galzin & Legendre, 1987; Ardisson *et al*., 1990; Tuomisto & Ruokolainen, 1994: Ecological application 12.6c; Tuomisto *et al*., 2003). When applied to *ecological succession*, chronological clustering corresponds to a well-defined *model*, in which succession proceeds by steps and the transitions between steps are rapid (see also Allen *et al*., 1977, on this topic). Broad-scale successional steps contain finer-scale steps, which may be identified using a finer analysis if finer-scale data are available. Chronological clustering takes into account the sampling sequence, imposing a constraint of temporal contiguity to the clustering activity.

The method also permits the elimination of *singletons* (in the game of *bridge*, a singleton is a card that is the only one of a suit in the hand of a player). Such singular observations often occur in ecological series. In nature, singletons are the result of

	Group 1		Group 2		
D	5	6	7	8	9
5	0				
6	0.2	0			(a)
7	0.4	0.7	0		
8	0.6	0.5	0.1	0	
9	0.7	0.8	0.3	0.6	0

	Group 1		Group 2		
D	5	6	7	8	9
5					
6	0				(b)
7	0	1			
8	1	0	0		
9	1	1	0	1	

Figure 12.22 Numerical example. (a) Distance matrix for two contiguous groups from a multidimensional time series (used also in Fig. 10.23). The lower half of the symmetric matrix is shown. (b) 50% of the distances, i.e. those with the highest values, are coded 1; the others are coded 0.

random fluctuations, migrations, or local changes in external forcing. In an aquatic system studied at a fixed location (Eulerian approach, Section 12.0), such changes may be due to temporary movements of water masses. Singletons may also result from improper sampling or inadequate preservation of specimens.

Hierarchical agglomerative clustering (Section 8.5) proceeds from an association matrix ($n \times n$) among the observations of the data series (length n), computed using an appropriately chosen similarity or distance coefficient (Chapter 7). Any method of agglomerative clustering may be used; Legendre *et al.* (1985) used intermediate linkage clustering (Subsection 8.5.3). The clustering algorithm is modified to include the contiguity constraint; Fig. 13.25 shows how a constraint of spatial or temporal contiguity can be introduced into any agglomerative clustering algorithm. Each clustering step is subjected to a permutation test (Subsection 1.2.2) before the fusion of two objects or groups is authorized.

Permutation test

Consider two adjacent groups of objects pertaining to some data series (Fig. 12.22). The first group ($n_1 = 2$) includes objects 5 and 6 and the second ($n_2 = 3$) contains objects 7, 8 and 9. Assume that an agglomerative clustering algorithm now proposes that these two groups are the next pair to join. Distances among the five objects are given in Fig. 12.22a. Before applying the permutation test of cluster fusion, the distances are divided in two groups: the 50% of the distances (5 in this example) that have the highest values are called "high distances" and are coded 1 (Fig. 12.22b) whereas the other 50% are called "low distances" and are coded 0. The test statistic is the number of high distances (h) in the between-group matrix (shaded area); $h = 4$ in this example. Under the null hypothesis, the objects in the two groups are drawn from the same statistical population and, consequently, it is only an artefact of the agglomerative clustering algorithm that they temporarily form two groups. If the null hypothesis is true, the number of high distances ($h = 4$) presently found in the between-group matrix should be comparable to that found among all possible

permutations of the five objects in two groups with $n_1 = 2$ and $n_2 = 3$ objects. If the null hypothesis is false and the two groups come from different statistical populations (i.e. different steps of the succession), the number of high distances presently found in the between-group matrix should be *higher* than most of the values found after permutation of the objects into two groups with $n_1 = 2$ and $n_2 = 3$ objects. This calls for a one-tailed test. After setting a significance level α, the permutations are performed and results that are higher than or equal to h are counted. The number of distinguishable combinations of the objects in two groups of sizes n_1 and n_2 is $(n_1 + n_2)!/(n_1!\, n_2!)$. If this number is not too large, all possible permutations can be examined; otherwise, permutations may be selected at random to form the reference distribution for significance testing. The number of permutations producing a result as large as or larger than h, divided by the number of permutations performed, gives an estimate of the probability p of observing the data under the null hypothesis.

- If $p > \alpha$, the null hypothesis is not rejected and the two groups are fused.
- If $p \leq \alpha$, the null hypothesis is rejected and fusion of the groups is prevented.

This test may actually be reformulated as a Mantel test (Section 10.5.1) between the matrix of recoded distances (Fig. 12.22b) and another matrix of the same size containing 1's in the among-group rectangle and 0's elsewhere.

Internal validation criterion

The above is not a proper test of significance because the alternative hypothesis (H_1: the two groups actually found by the clustering method differ) is not independent of the data that are used to perform the test; it comes from the data through the agglomerative clustering algorithm. So this is actually an internal validation clustering criterion (Section 8.13). Legendre *et al*. (1985) have shown, however, that this criterion has a correct probability of type I error; when testing on randomly generated data (Monte Carlo simulations) at significance level α, the null hypothesis was rejected in a proportion of the cases approximately equal to α.

Resolution

Significance level α used as the criterion for cluster fusion determines how easy it is to reject the null hypothesis. When α is small (close to 0), the null hypothesis is almost never rejected and only the sharpest discontinuities in the time or space series are identified. Increasing the value of α actually makes it easier to reject the null hypothesis, so that more groups are formed; the resulting groups are thus smaller and bring out more discontinuities in the data series. So, changing the value of α actually changes the resolution of the clustering results.

Singleton

A singleton is defined as a single observation whose fusion has been rejected with the groups located to its right and left in the series. When the test leads to the discovery of a singleton, it is temporarily removed from the series and the clustering procedure is started again from the beginning. This is done because the presence of a singleton can disturb the whole clustering geometry, as a result of the contiguity constraint.

The end result of chronological clustering is a *nonhierarchical partition* of the series into nonoverlapping homogeneous groups. Within the context of ecological succession, these groups correspond to the steps of a succession. *A posteriori* tests are used to assess the relationships between distant groups along the series as well as the origin of singletons. Plotting the clusters of observations onto an ordination diagram in reduced space may help in the overall interpretation of the results.

Legendre (1987b) showed that time-constrained clustering possesses some interesting properties. On the one hand, applying a constraint of spatial or temporal contiguity to an agglomerative clustering procedure forces different clustering methods to produce approximately the same results; without the constraint, the methods may lead to very different clustering results (Chapter 8), except when the spatial or temporal structure of the data (patchiness, gradient: Section 13.0) is very strong. Using autocorrelated simulated data series, he also showed that, if patches do exist in the data, constrained clustering always recovers a larger fraction of the structure than the unconstrained equivalent.

Constrained clustering along a time or spatial sampling axis can also be done by a more general form of constrained hierarchical clustering described in Subsection 13.3.2; see function ***constrained.clust()*** in Section 12.8.

Ecological application 12.6b

In May 1977, the Société d'Énergie de la Baie James impounded a small reservoir (ca. 7 km^2), called Desaulniers, in Northern Québec (77°32' W, 53°36' N). Ecological changes occurring during the operation were carefully monitored in order to use them to forecast the changes that would take place upon impoundment of much larger hydroelectric reservoirs in the same region. Several sampling sites were visited before and after the flooding. Effects of flooding on the zooplankton community of the deepest site (max. depth: 13 m), located ca. 800 m from the dam, were studied by Legendre *et al*. (1985) using chronological clustering. Before flooding, the site was located in a riverbed and only zooplankton drifting from lakes located upstream was found there (i.e. there was no zooplankton community indigenous to the river). Changes observed are thus an example of primary succession.

After logarithmic normalization of the data (eq. 1.14), the Canberra metric (D_{10}, eq. 7.49) was used to compute distances among all pairs of the 47 observations. Homogeneous groups of observations were identified along the data series, using a time-constrained algorithm for intermediate linkage clustering (Subsection 8.5.3) and the permutation test of cluster fusion described above. Results of chronological clustering are shown in Fig. 12.23 for different levels of resolution α. Plotting the groups of observations from Fig. 12.23, for $\alpha = 0.25$, on an ordination diagram obtained by nonmetric multidimensional scaling (Fig. 12.24), led to the following conclusions concerning changes in the zooplankton community. In 1976, as mentioned above, zooplankton was drifting randomly from small lakes located upstream. This was evidenced by low species numbers and highly fluctuating evenness (eq. 6.45), which indicated that no stable community was present. After impoundment of the reservoir, the community departed rapidly from the river status (Fig. 12.24) and formed a fairly well-developed assemblage, with 13 to 20 species in the summer of 1977, despite large chemical and water-level fluctuations. After the autumn overturn and during the 1977-1978 winter period, the

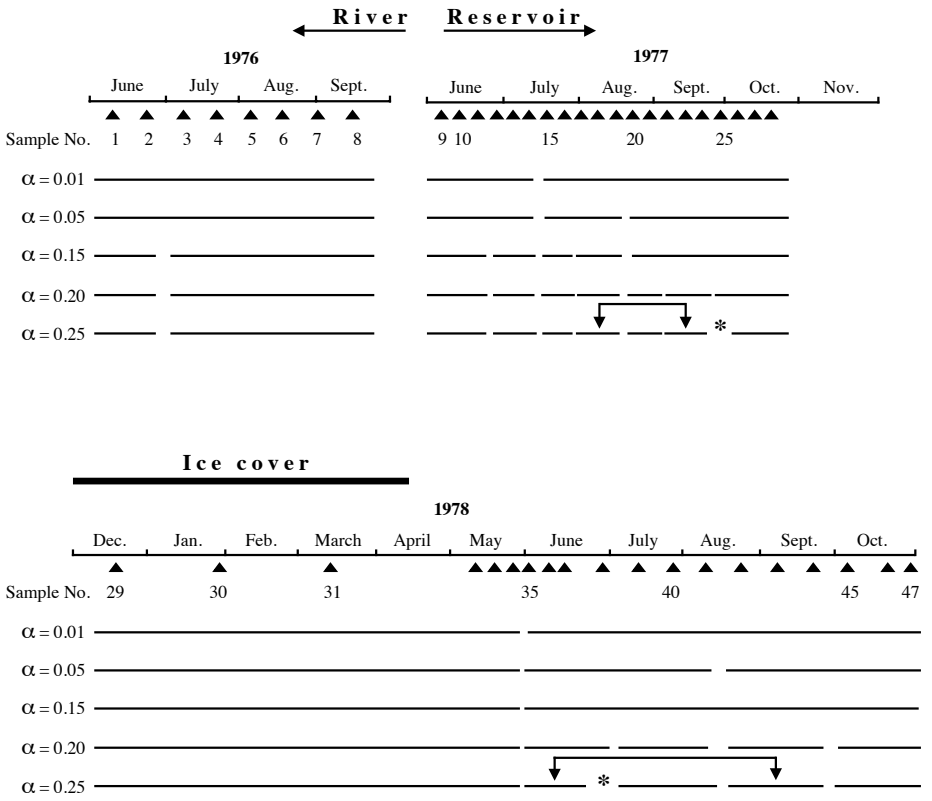


Figure 12.23 Chronological clustering: zooplankton time series. Results for different levels of resolution (α). For α = 0.25, the double arrows identify *a posteriori* tests with probabilities of fusion larger than α. Asterisks (*) identify singletons. Modified from Legendre *et al*. (1985).

community moved away from the previous summer's status. When spring came (observation 35), the community had reached a zone of the multidimensional scaling plane quite distinct from that occupied in summer 1977. Zooplankton was then completely dominated by rotifers, which increased from 70 to 87% in numbers and from 18 to 23% in biomass between 1977 and 1978, with a corresponding decrease in crustaceans, while the physical and chemical conditions had stabilized (Pinel-Alloul *et al*., 1982). When the succession was interrupted by the 1978 autumn overturn, the last group in the series (observations 45-47) was found (Fig. 12.23) near the position of the previous winter's observations (29-34), indicating that the following year's observations might resemble the 1978 succession.

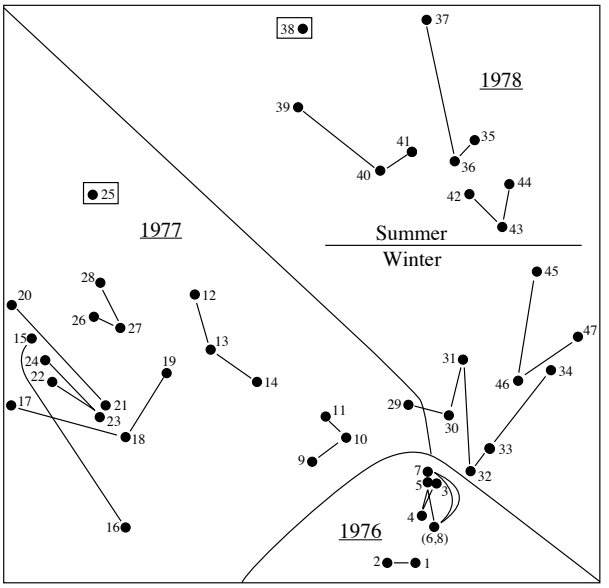


Figure 12.24 Chronological clustering: zooplankton time series. Nonmetric multidimensional scaling plot showing groups of observations from Fig. 12.23, for $\alpha = 0.25$. The groups are the sets of observations that are connected by lines materializing the sampling sequence. Objects in boxes are singletons. From Legendre *et al*. (1985). The regions of the graph delimited by envelopes correspond to sampling years.

Ecological application 12.6c

Tuomisto & Ruokolainen (1994) studied species assemblages of *Pteridophyta* (ferns; 40 species in the study) and *Melastomataceae* (a family of shrubs, vines, and small trees restricted to the American tropics; 22 species in the study) along two spatial transects (replicates) in a non-flooded area of the Amazonian rain forest in Peru, covering an edaphic (i.e. soil-related) and topographic gradient from clay soil on level ground, to quarzitic sand on a hill top. The two 700-m-long and 5-m-wide, parallel transects were 50 m apart. Chronological clustering was applied to the edaphic and floristic variables separately, using different similarity coefficients and three levels of resolution (parameter α). In all cases, the transects could be divided into distinct sections; the results of constrained clustering were more readily interpretable than the unconstrained equivalent. The groups of plants selected proved adequate for the rapid assessment of changes in the floristic composition of the rain forest.

Ecological application 12.6d

Tuomisto *et al*. (2003) studied the community structure of *Pteridophyta* (ferns) and *Melastomataceae*, the same groups as in Ecological application 12.6c, along a 43-km long transect in the Amazonian rain forest in Peru. They segmented the series of pteridophytes and

Melastomataceae data into groups using chronological clustering. They also used chronological clustering to partition a data series of spectral reflectance characteristics of the forest, extracted from a Landsat TM satellite image. The chronological clustering results were fairly consistent; the authors recognize eight groups of sites, which were also related to topography and soil characteristics. *Pteridophyta* and *Melastomataceae* indicator species of these groups of sites were then identified using the *INDVAL* index (Subsection 8.9.3). The results supported the hypothesis that species segregate edaphically at the landscape scale within the rain forest.

12.7 Box-Jenkins models

Forecasting

Objective 6 of time series analysis in ecology (Section 12.1) is to *forecast* future values. The Preface explained that ecological modelling is not, as such, within the scope of numerical ecology. In ecological studies, however, *Box-Jenkins modelling* is often conducted together with other forms of series analysis; this is why it is briefly presented here. This type of technique has already been mentioned in the context of maximum entropy spectral analysis (MESA, Section 12.5.3). The present section summarizes the principles that underlie the approach. Interested readers may refer to Box & Jenkins (1976), Cryer (1986), and Bowerman & O'Connell (1987) for the theory and to user's manuals of computer packages and R functions for actual implementation of the method.

MA model

Stochastic linear models (processes) described here are based on the idea that, in a series where data within a small window are strongly interrelated, the observed values are generated by a number of "shocks" a_t. These shocks are independent of each other and their distribution is purely random (mean zero and variance s_a^2). Such a series $(a_t, a_{t-1}, a_{t-2}, \ldots)$ is called *white noise*. In the *moving average* (*MA*) *model*, each observations in the series ($\tilde{y}_t = y_t - \bar{y}$, i.e. the data are centred on the mean $\bar{y}$ of the series) can be represented as a weighted sum of the values of process a:

$$\tilde{y}_t = a_t - (\theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}) \qquad \textbf{(12.31)}$$

where θ are the weights and q is the *order* of the model. The name *moving average* for this process comes from the fact that eq. 12.31 is somewhat similar to that of the moving average (see the right-hand column of Table 12.4). The weights θ are estimated by numerical iteration, using techniques that are described in the above references and available in computer packages and R functions.

When the above model does not fit the series adequately (see below), another possibility is to represent an observation by a weighted sum of the q previous observations plus a random shock:

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \ldots + \phi_q \tilde{y}_{t-q} + a_t$$

AR model

This is the *autoregressive model* (*AR*, or *all-pole* model) already described in eq. 12.29. In this model (of *order* q), q successive terms of the series are used to

forecast term $(q + 1)$, with error a_t. When estimating the autocorrelation coefficients ϕ by least squares, it is easy to compute residual errors $a_t = y_t - \tilde{y}_t$. Residual errors, as specified above for all Box-Jenkins models, must be independent of one another; this implies that a correlogram of the series of residuals a_t should display no significant value. The residuals must also be normally distributed.

ARMA model

Combining the above two models gives the *autoregressive-moving average model* (*ARMA model*), whose general form is:

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \ldots + \phi_q \tilde{y}_{t-q} + a_t - (\theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}) \quad \textbf{(12.32)}$$

An important advantage of ARMA models is that they can be fitted to data series using a small number of parameters (i.e. the coefficients ϕ and θ). However, such models may only be estimated for strictly *stationary* series (Sections 12.1 and 12.2).

One approach described in Section 12.2 for extracting the trend from a series is the *variate difference method*. In the computation, each value y_t is replaced by $(y_t - y_{t-T})$ where T is the period of the trend:

$$\tilde{y}_t = y_t - y_{t-T} \quad \textbf{(12.33)}$$

ARIMA model

Since $\tilde{y}_t$ results from a *difference*, y_t is called the *integrated form* of $\tilde{y}_t$. When an ARMA model is applied to a series of values computed with eq. 12.33, it is called an *autoregressive-integrated-moving average model* (*ARIMA model*).

Box-Jenkins analysis normally proceeds in four steps. (1) Identification of the *type of model* to be fitted to the data series (i.e. MA, AR, ARMA, or ARIMA). Even though Box & Jenkins (1976) described some statistical properties of the series (e.g. shape of the autocorrelation) that may guide this choice, identification of the proper model remains a somewhat intuitive step (e.g. Ibanez, 1982). (2) Estimation of the *parameters* of the model. For each case, various methods are generally available, so that one is confronted with a choice. (3) The *residuals* must be independent and normally distributed. If not, either the model is not adequate for the data series, or the parameters were not properly estimated. In such a case, step (2) can be repeated with a different method or, if this does not improve the residuals, a different model must be chosen at step (1). Steps (1) through (3) may be repeated as many times as necessary to obtain a good fit. The procedure of identification of the appropriate model is therefore iterative. (4) Using the model, values can be *forecasted* beyond the last observation.

It may happen that the data series is under external influences, so that the models described above cannot be used as such. For example, in the usual ARIMA model, the state of the series at time t is a function of the previous q observations ($\tilde{y}$) and of the random errors (a). In order to account for the additional effect of external variables, some computer programs allow the inclusion of a *transfer function* into the model (if the external forcing variable is also a random variable) and/or an *intervention component* (if the external variable is binary and not random). It is possible to extend

the forecasting to *multidimensional* data series. References to conduct the analysis are Whittle (1963) and Jones (1964).

It is important to remember that the models discussed here are *forecasting* and not *predictive* models. Indeed, the purpose of Box-Jenkins modelling is to *forecast* values of the series beyond the last observation, using the preceding data. Such forecasting is only valid as long as the environmental conditions that characterize the population under study (demographic rates, migrations, etc.) as well as the anthropogenic effects (exploitation methods, pollution, etc.) remain essentially the same. In order to *predict* with some certainty the fate of the series, causal relationships should be determined and modelled; for example, between the observed numbers of organisms, on the one hand, and the main environmental conditions, population characteristics, or/and anthropogenic factors, on the other. This requires extensive knowledge of the system under study. Forecasting models often prove quite useful in ecology, but one must be careful not to use them beyond their limits.

Ecological application 12.7

Boudreault *et al*. (1977) tried to forecast lobster landings in Îles-de-la-Madeleine (Gulf of St. Lawrence, Québec), using various methods of series analysis. In a first step, they found that an *autoregressive model* (of order 1) accounted for ca. 40% of the variance in the series of landings. This relatively low percentage could be explained by the fact that observations in the series were not very homogeneous. In a second step, external physical variables were added to the model and the data were analysed using *regression on principal components* (Section 10.3). The two external variables were: water temperature in December, 8.5 years before the fishing season, and average winter temperature 3.5 years before. This increased to 90% the variance explained by the model. Lobster landings in a given year would thus depend on: the available stock (autocorrelated to landings during the previous year), the influence of water temperature on larval survival (lobster *Homarus americanus* around Îles-de-la-Madeleine reach commercial size when ca. 8 years old), and the influence of water temperature at the time the animals reached sexual maturity (at the age of ca. 5 years).

12.8 Software

Procedures available in commercial statistical packages are not reviewed here. The R language offers functions for the methods described in Chapter 12.

1. Time series objects. — Function ***ts()*** of STATS creates a time-series object identified to class "ts". ***plot.ts()*** plots a graph for such an object, ***ts.plot()*** plots several time series in a common plot. ***ts.union()*** binds two or more time series into a single R object.

2. Trend extraction. — Function ***lm()*** in STATS is used to detrend data, i.e. extract a linear or polynomial trend and compute residuals.

3. Periodic variability: correlogram. — Function ***acf()*** in STATS computes spatial autocovariance and autocorrelation; ***plot.acf()*** plots confidence intervals under either a white noise or a MA model. ***ccf()*** computes cross-covariance and cross-correlation. For spatial transects or time series with irregular lags, correlograms can be computed using function ***sp.correlogram()*** of package SPDEP; a constant must be written in the second column in the file of geographic coordinates used to create the list of neighbours.

4. Periodic variability: periodogram. — ***buysbal()*** in PASTECS constructs Buys-Ballot tables from time series. ***periodograph()***[*] computes the contingency periodogram (Subsection 12.4.2). ***spec.pgram()*** in STATS estimates the spectral density of a series by a smoothed Schuster periodogram. ***cpgram()*** plots a cumulative periodogram.

5. Periodic variability: spectral analysis. — Function ***spectrum()*** in STATS estimates the spectral density of a time series. ***spec.ar()*** fits an AR model to a time series and computes the spectral density of the fitted model.

6. Wavelet analysis. — Function ***dwt()*** of WAVESLIM is used to compute wavelet analysis for data series, and ***dwt.2d()*** for two-dimensional data[†]. Package WMTSA contains other wavelet methods for time series analysis.

7. Detection of discontinuities in multivariate series. — Function ***chclust()*** of package RIOJA, developed for palaeoecological reconstruction, performs constrained hierarchical clustering from a distance matrix, with clusters constrained by the order of the sampling units in the data file. The method is applicable to temporal or spatial multivariate series, such as sediment core data. Multivariate regression tree analysis (MRT) can also be used for constrained clustering for temporal or spatial multivariate data series, as shown in Subsection 12.6.4. Function ***constrained.clust()*** of package CONST.CLUST[*] carries out constrained hierarchical clustering along a time or spatial series, or on a geographic surface (Section 13.3.2), with cross-validation of the results. Chronological clustering (Section 12.6.5) is implemented in function ***chrono*** of THE R PACKAGE[*] for mainframe computers and Mac OS Classic. This program has not been rewritten yet for the R language.

8. Box-Jenkins models. — Function ***ar()*** in STATS fits an autoregressive model to a univariate or multivariate time series; ***arima()*** fits an ARIMA model to a univariate time series; ***ARMAacf()*** computes the theoretical autocorrelation function for an ARMA process.

9. Miscellaneous methods. — Function ***turnogram()*** in PASTECS computes and plots turnograms; ***turpoints()*** analyses turning points (Section 12.1) and tests the randomness of series.

---

[*] Available on the Web page http://numericalecology.com/rcode.

[†] An introduction and R code for wavelet analysis using WAVESLIM are found on the Web page https://sites.google.com/site/patrickmajames/stuff.