CHAPTER 5

Time-series Analysis Methods

The advent of high-density storage devices and long-term mooring capability has enabled oceanographers to collect long time series of oceanic and meteorological data. Similarly, the use of rapid-response sensors on moving platforms has made it possible to generate snapshots of spatial variability over extensive distances. Time-series data are collected from moored instrument arrays or by repeated measurements at the same location using ships, satellites, or other instrumented packages. Quasi-synoptic spatial data are obtained from ships, manned-submersibles, remotely operated vehicles (ROVs), autonomous underwater vehicles (AUVs), satellites, and satellite-tracked drifters.

As discussed in Chapters 3 and 4, the first stage of data analysis following data verification and editing usually involves estimates of arithmetic means, variances, correlation coefficients, and other sample-derived statistical quantities. These quantities tell us how well our sensors are performing and help characterize the observed oceanographic variability. However, general statistical quantities provide little insight into the different types of signals that are blended together to make the recorded data. The purpose of this chapter is to present methodologies that examine data series in terms of their frequency content. With the availability of modern high-speed computers, frequency-domain analysis has become much more central to our ability to decipher the cause and effect of oceanic change. The introduction of fast Fourier transform (FFT) techniques in the 1960s further aided the application of frequency-domain analysis methods in oceanography.

5.1 BASIC CONCEPTS

For historical reasons, the analysis of sequential data is known as *time series analysis*. As a form of data manipulation, it has been richly developed for a wide assortment of applications. While we present some of the latest techniques, the emphasis of this chapter will be on those "tried and proven" methods most widely accepted by the general oceanographic community. Even these established methods are commonly misunderstood and incorrectly applied. Where appropriate, references to other texts will be given for those interested in a more thorough description of analysis techniques. As with previous texts, the term "time series" will be applied to both temporal and spatial data series; methods which apply in the time domain also apply in the space domain. Similarly, the terms *frequency domain* and *wavenumber domain* (the

formal transforms of the time and spatial series, respectively) are used interchangeably.

A basic purpose of time series analysis methods is to define the variability of a data series in terms of dominant periodic functions. We also want to know the "shape" of the spectra. Of all oceanic phenomena, the barotropic astronomically forced tides most closely exhibit deterministic and stationary periodic behavior, making them the most readily predictable motions in the sea. In coastal waters, tidal observations over a period as short as one month can be used to predict local tidal elevations with a high degree of accuracy. Where accurate specification of the boundary conditions is possible, a reasonably good hydrodynamic numerical model that has been calibrated against observations can reproduce the regional tide heights to an accuracy of a few centimeters. Tidal currents are less easily predicted because of the complexities introduced by stratification, nonlinear interactions, and basin topography. Although baroclinic (internal) tides generated over abrupt topography in a stratified ocean have little impact on surface elevations, they can lead to strong baroclinic currents. These currents are generally stochastic (i.e. nondeterministic) and hence only predictable in a statistical sense.

Surface gravity waves are periodic and quasi-linear oceanic features but are generally stochastic due to inadequate knowledge of the surface wind fields, the air–sea momentum transfer, and oceanic boundary conditions. Refraction induced by wave–current interactions can be important but difficult to determine. Other oceanic phenomena such as coastal-trapped waves and near-inertial oscillations have marked periodic signatures but are intermittent because of the vagaries of the forcing mechanisms and changes in oceanic conditions along the direction of propagation. Other less obvious regular behavior can be found in observed time and space records. For instance, oceanic variability at the low-frequency end of the spectrum is dominated by fluctuations at the annual to decadal periods, consistent with baroclinic Rossby waves and short-term climate change, while that at the ultra-low frequencies is dominated by ice-age climate scale variations associated with Milankovitch-type processes (changes in the caloric summer insolation at the top of the atmosphere arising from changes in the earth's orbital eccentricity, and tilt and precision of its rotation axis).

Common sense should always be a key element in any time-series analysis. Attempts to use analytical techniques to find "hidden" signals in a time series often are not very convincing, especially if the expected signal is buried in the measurement noise. Because noise is always present in real data, it should be clear that, for accurate resolution of periodic behavior, data series should span at least a few repeat cycles of the time scale of interest, even for stationary processes. Thus, a day-long record of hourly values will not fully describe the diurnal cycle in the tide nor will a 12-month series of monthly values fully define the annual cycle of sea surface temperature. For these short records, modern spectral analysis methods can help us pin-point the peak frequencies. As we noted in Chapter 1, a fundamental limitation to resolving time-series fluctuations is given by the "sampling theorem" which states that the highest detectable frequency or wavenumber (the Nyquist frequency or wavenumber) is determined by the interval between the data points. For example, the highest frequency that we can resolve by an hourly time series is one cycle per 2 h, or one cycle per $2\Delta t$, where Δt is the interval of time between points in the series.

For the most part, we fit series of well-known functions to the data in order to transform from the time domain to the frequency domain. As with the coefficients of

the sine and cosine functions used in Fourier analysis, we generally assume that the functions have slowly varying amplitudes and phases, where "slowly" means that coefficients change little over the length of the record. Other linear combinations of orthogonal functions with similar limitations on the coefficients can be used to describe the series. However, the trigonometric functions are unique in that uniformly spaced samples covering an integer number of periods of the function form orthogonal sequences. Arbitrary orthogonal functions, with a similar sampling scheme, do not necessarily form orthogonal sequences. Another advantage of using common functions in any analysis is that the behavior of these functions is well understood and can be used to simplify the description of the data series in the frequency or wavenumber domain. In this chapter, we consider time series to consist of periodic and aperiodic components superimposed on a secular (long-term) trend and uncorrelated random noise. Fourier analysis and spectral analysis are among the tools used to characterize oceanic processes. Determination of the Fourier components of a time series can be used to determine a *periodogram* which can then be used to define the spectral density (*spectrum*) of the time series. However, the periodogram is not the only way to get at the spectral energy density. For example, prior to the introduction of the fast Fourier transform (FFT), the common method for calculating spectra was through the Fourier transform of the autocorrelation function. More modern spectral analysis methods involve autoregressive spectral analysis (including use of maximum entropy techniques), wavelet transforms, and fractal analysis.

5.2 STOCHASTIC PROCESSES AND STATIONARITY

A common goal of most time-series analysis is to separate deterministic periodic oscillations in the data from random and aperiodic fluctuations associated with unresolved background noise (unwanted geophysical variability) or with instrument error. It is worth recalling that time-series analyses are typically statistical procedures in which data series are regarded as subsets of a stochastic process. A simple example of a stochastic process is one generated by a linear operation on a purely random variable. For example, the function $x(t_i) = 0.5x(t_{i-1}) + \varepsilon(t_i)$, $i = 1, 2, \ldots$, for which $x(t_0) = 0$, say, is a linear random process provided that the fluctuations $\varepsilon(t_i)$ are statistically independent. Stochastic processes are classified as either discrete or continuous. A continuous ("analog") process is defined for all time steps while a discrete ("digital") process is defined only at a finite number of points. The data series can be scalar (univariate series) or a series of vectors (multivariate series). While we will deal with discrete data, we assume that the underlying process is continuous.

If we regard each data series as a realization of a stochastic process, each series contains an infinite ensemble of data having the same basic physical properties. Since a particular data series is a sample of a stochastic process, we can apply the same kind of statistical arguments to our data series as we did to individual random variables. Thus, we will be making statistical probability statements about the results of frequency transformations of our data series. This fact is important to remember since there is a great temptation to regard transformed values as inherently independent data points. Since many data collected in time or space are highly correlated because of the presence of low-frequency, nearly deterministic components, such as long-

period tides and the seasonal cycle, standard statistical methods do not really apply. Contrary to the requirements of stochastic theory, the values are not statistically independent. "What constitutes the ensemble of a possible time series in any given situation is dictated by good scientific judgment and not by purely statistical matters" (Jenkins and Watts, 1968). A good example of this problem is presented by Chelton (1982) who showed that the high correlation between the integrated transport through Drake Passage in the Southern Ocean and the circumpolar-averaged zonal wind stress "may largely be due to the presence of a strong semi-annual signal in both time series." A strong statistical correlation does not necessarily mean there is a cause and effect relationship between the variables.

As implied by the previous section, the properties of a stochastic process generally are time dependent and the value $y(t)$ at any time, t, depends on the time elapsed since the process started. A simplifying assumption is that the series has reached a steady state or equilibrium in the sense that the statistical properties of the series are independent of absolute time. A minimum requirement for this condition is that the probability density function (PDF) is independent of time. Therefore, a stationary time series has constant mean, μ, and variance, σ^2. Another consequence of this equilibrium state is that the joint PDF depends only on the time difference $t_1 - t_2 = \tau$ and not on absolute times, t_1 and t_2. The term *ergodic* is commonly used in association with stochastic processes for which time averages can be used in place of ensemble averages (see Chapter 3). That is, we can average over "chunks" of a time series to get the mean, standard deviation, and other statistical quantities rather than having to produce repeated realizations of the time series. Any formalism involving ensemble averaging is of little value as the analyst rarely has an ensemble at his disposal and typically must deal with a single realization. We need the ergodic theorem to enable us to use time averages in place of ensemble averages.

5.3 CORRELATION FUNCTIONS

Discrete or continuous random time series, $y(t)$, have a number of fundamental statistical properties that help characterize the variability of the series and make it easily possible to compare one time series against another. However, these statistical measures also contain less information than the original time series and, except in special cases, knowledge of the these properties is insufficient to reconstruct the time series.

(1) *Mean and variance.* If y is a stochastic time series consisting of N values $y(t_i) = y_i$ measured at discrete times t_i $\{t_1, t_2, \ldots, t_N\}$, the true mean value μ for the series can be estimated by

$$\mu \equiv E[y(t)] = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{5.3.1}$$

where $E[y(t)]$ is the expected value and $E[|y(t)|] < \infty$ for all t. The estimated mean value is not necessarily constant in time; different segments of a time series can have different mean values if the series is nonstationary. If $E[y^2(t)] < \infty$ for all t, an

estimate of the true variance function is given by

$$\sigma^2 \equiv E[\{y(t) - \mu\}^2] = \frac{1}{N}\sum_{i=1}^{N}[y_i - \bar{y}]^2 \tag{5.3.2}$$

The positive square root of the variance is the standard deviation, σ, or root-mean-square (RMS) value. See Chapter 3 for further discussion on the mean and variance.

(2) *Covariance and correlation functions*: These terms are used to describe the covariability of given time series as functions of two different times, $t_1 = t$ and $t_2 = t + \tau$, where τ is the lag time. If the process is *stationary* (unchanging statistically with time) as we normally assume, then absolute time is irrelevant and the covariance functions depend only on τ.

Although the terms "covariance function" and "correlation function" are often used interchangeably in the literature, there is a fundamental difference between them. Specifically, covariance functions are derived from data series following removal of the true mean value, μ, which we typically approximate using the sample mean, $\overline{y(t)}$. Correlation functions use the "raw" data series before removal of the mean. The confusion arises because most analysts automatically remove the mean from any time series with which they are dealing. To further add to the confusion, many oceanographers define correlation as the covariance normalized by the variance.

For a stationary process, the *autocovariance function*, C_{yy}, which is based on lagged correlation of a function with itself, is estimated by

$$\begin{aligned} C_{yy}(\tau) &\equiv E[\{y(t) - \mu\}\{y(t+\tau) - \mu\}] \\ &= \frac{1}{N-k}\sum_{i=1}^{N-k}[y_i - \bar{y}][y_{i+k} - \bar{y}] \end{aligned} \tag{5.3.3}$$

where $\tau = \tau_k = k\Delta t$ $(k = 0, \ldots, M)$ is the lag time for k sampling time increments, Δt, and $M \ll N$. The corresponding expression for the *autocorrelation function* R_{yy} is

$$\begin{aligned} R_{yy}(\tau) &\equiv E[y(t)y(t+\tau)] \\ &= \frac{1}{N-k}\sum_{i=1}^{N-k}(y_i y_{i+k}) \end{aligned} \tag{5.3.4}$$

At zero lag $(\tau = 0)$

$$C_{yy}(0) = \sigma^2 = R_{yy}(0) - \mu^2 \tag{5.3.5}$$

where we must be careful to define σ^2 in equation (5.3.2) in terms of the normalization factor $1/N$ rather than $1/(N-1)$ (see Chapter 3). From the above definitions, we find

$$C_{yy}(\tau) = C_{yy}(-\tau);\ R_{yy}(\tau) = R_{yy}(-\tau) \tag{5.3.6}$$

indicating that the autocovariance and autocorrelation functions are symmetric with respect to the time lag τ.

The autocovariance function can be normalized using the variance (5.3.2) to yield the normalized autocovariance function

$$\rho_{yy}(\tau) = \frac{C_{yy}(\tau)}{\sigma^2} \tag{5.3.7}$$

(Note: some oceanographers call (5.3.7) the autocorrelation function.)

The basic properties of the normalized autocovariance function are:

(a) $\rho_{yy}(\tau) = 1$, for $\tau = 0$;
(b) $\rho_{yy}(\tau) = \rho_{yy}(-\tau)$, for all τ;
(c) $|\rho_{yy}(\tau)| \leq 1$, for all τ;
(d) If the stochastic process is continuous, then $\rho_{yy}(\tau)$, must be a continuous function of τ.

If we now replace one of the $y(t)$ in the above relations with another function $x(t)$, we obtain the *cross-covariance function*

$$\begin{aligned} C_{xy}(\tau) &\equiv E[\{y(t) - \mu_y\}\{x(t+\tau) - \mu_x\}] \\ &= \frac{1}{N-k}\sum_{i=1}^{N-k}[y_i - \bar{y}][x_{i+k} - \bar{x}] \end{aligned} \tag{5.3.8}$$

and the *cross-correlation function*

$$\begin{aligned} R_{xy}(\tau) &\equiv E[y(t)x(t+\tau)] \\ &= \frac{1}{N-k}\sum_{i=1}^{N-k} y_i x_{i+k} \end{aligned} \tag{5.3.9}$$

The normalized cross-covariance function (or *correlation coefficient function*) for a stationary process is

$$\rho_{xy} \equiv \frac{C_{xy}(\tau)}{\sigma_x \sigma_y} \tag{5.3.10}$$

Here, $y(t)$ could be the longshore component of daily mean wind stress and $x(t)$ the daily mean sea level elevation at the coast. Typically, sea level lags the longshore wind stress by one to two days.

One should be careful interpreting covariance and correlation estimates made for large lags. Problems arise if low-frequency components are present in the data since the averaging inherent in these functions becomes based on fewer and fewer samples and loses its statistical reliability as the lag increases. For example, at lag $\tau = 0.1T$ (i.e. 10% of the length of the time series) there are roughly 10 independent cycles of any variability on a time scale, $T_{0.1} = 0.1T$, while at lags of $0.5T$ there are only about two independent estimates of the time scale $T_{0.5}$. In many cases, low-frequency components in geophysical time series make it pointless to push the lag times much beyond 10–20% of the data series. Some authors argue that division by N rather than by $N - k$ reduces the bias at large lags. Although this is certainly true ($N \gg N - k$ at large lags), it doesn't mean that the result has anything to do with reality. In essence, neither of these estimators are optimal. Ideally one should write down the likelihood function of the observed time series, if it exists. Differentiation of this likelihood function would then give a set of equations for the maximum likelihood estimates of

the autocovariance function. Unfortunately, the derivatives are in general untraceable and one must work with estimators given above. Results for this section are summarized as follows:

(a) Estimators with divisors $T = N\Delta t$ usually have smaller mean square errors than those based on $T - \tau$; also, those based on 1/T are positive definite while those based on $1/(T - \tau)$ may not be.
(b) Some form of correction for low-frequency trends is required. In simple cases, one can simply remove a mean value while in others the trend can be removed. Trend removal must be done carefully so that erroneous data are not introduced into the time series during the subtraction of the trend.
(c) There will be strong correlations between values in the autocorrelation function if the correlation in the original series was moderately strong; the autocorrelation function, which can be regarded as a new time series derived from $y(t)$, will, in general, be more strongly correlated than the original series.
(d) Due to the correlation in (c), the autocorrelation function may fail to dampen according to expectations; this will increase the basic length scale in the function.
(e) Correlation is a relative measure only.

In addition to its direct application to time-series analysis, the autocorrelation function was critical to the development of early spectral analysis techniques. Although modern methods typically calculate spectral density distributions directly from the Fourier transforms of the data series, earlier methods determined spectral estimates from the Fourier transform of the autocorrelation function. An important milestone in time-series analysis was the proof by N. Wiener and A. Khinchin in the 1930s that the correlation functions are related to the spectral density functions function through Fourier transform relationships. According to the Wiener–Khinchin relations, the autospectrum of a time series is the Fourier transform of its autocorrelation function.

(3) *Analytical correlation/covariance functions*: The autocorrelation function of a zero-mean random process $\varepsilon(t)$ ("white noise") can be written as

$$R_{\varepsilon\varepsilon}(\tau) = \sigma_\varepsilon^2 \rho_{\varepsilon\varepsilon}(\tau) = \sigma_\varepsilon^2 \delta(\tau) \tag{5.3.11}$$

where $\delta(\tau)$ is the Dirac delta function. In this example, σ_ε^2 is the variance of the data series. Another useful function is the cross-correlation between the time-lagged stationary signal $y(t) = \alpha x(t - \tau) + \varepsilon$ and the original signal $x(t)$. For constant α

$$R_{xy}(\tau) = \alpha R_{xx}(\tau - \tau_o) + \sigma_\varepsilon^2 \tag{5.3.12a}$$

which, for low noise, has a peak value

$$R_{xy}(\tau_o) = \alpha R_{xx}(0) = \alpha \sigma_x^2 \tag{5.3.12b}$$

Functions of the type (5.3.12) have direct use in ocean acoustics where the time lag, τ_o, at the peak of the zero-mean autocorrelation function can be related to the phase speed c and distance of travel d of the transmitted signal $x(t)$ through the relation $\tau_o = d/c$. It is through calculations of this type that modern acoustic Doppler current meters (ADCMs) and scintillation flow meters determine oceanic currents. In the case

of ADCMs, knowing τ_o and d gives the speed c and hence the change of the acoustic signal by the currents during the two-way travel time of the signal. Scintillation meters measure the delay τ_o for acoustic signals sent between a transmitter–receiver pair along two parallel acoustic paths separated by a distance d. The relation $\tau_o = d/v$ then gives the mean flow speed, v, normal to the direction of the acoustic path. Sending the signals both ways in the transmitter–receiver pairs gets around the problem of knowing the sound speed c in detail.

Although the calculation of autocorrelation and autocovariance functions is fairly straightforward, one must be very careful in interpreting the resulting values. For example, a stochastic process is said to be Gaussian (or normal) if the multivariate probability density function is normal. Then the process is completely described by its mean, variance, and autocovariance function. However, there is a class of nonGaussian processes which have the same normalized autocovariance function, ρ, as a given normal process. Consider the linear system

$$\tau_o \frac{dy}{dt} + y(t) = z(t) \tag{5.3.13}$$

where $z(t)$ is white-noise input and $y(t)$ is the output. Here, $y(t)$ is called a "first-order autoregressive process" which has the normalized autocorrelation function

$$\rho_{yy}(\tau) = e^{-|\tau|/\tau_o} \tag{5.3.14}$$

Thus, if the input to the first-order system has a normal distribution then by an extension of the central limit theorem it may be shown that the output is normal and is completely specified by the autocorrelation function.

Another process with an exponential autocorrelation function which differs greatly from the normal process is called the *random telegraph signal* (Figure 5.3.1). Alpha particles from a radioactive source are used to trigger a flip-flop between +1 and −1. Assuming the process was started at $t = -\infty$ we can derive the normalized autocorrelation function as

$$\rho_{yy}(\tau) = e^{-2\lambda|\tau|} \tag{5.3.15}$$

If $\lambda = 1/2\tau_o$ then this is the same as the autocorrelation function of a normal process, which is characteristically different from the flip-flop time series. Again, one must be careful when interpreting autocorrelation functions.

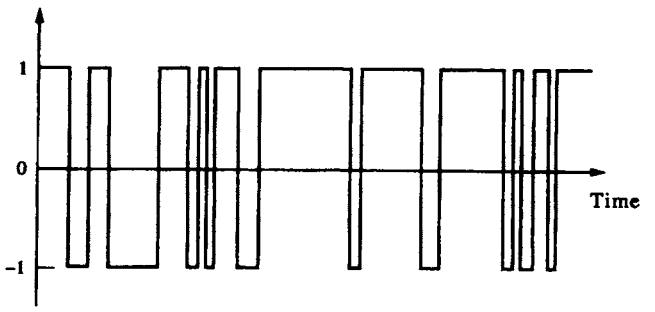


Figure 5.3.1. A realization of a random telegraph signal with digital amplitudes of ±1 as a function of time.

Table 5.1. Acoustic backscatter anomaly (decibels) measured in bin #1 from two adjacent transducers on a towed 150 kHz ADCP. The data cover a depth range of 75–230 m at increments of 5 m (32 values). The two vertical profiles are separated horizontally by a distance of 3.5 m. The means have not been removed from the data

Beam	75 m	80	85	90	95	100	105	110	115	120
1	11.56	0.67	−8.33	−9.82	−13.91	−18.00	3.67	−2.00	−12.29	−13.71
2	14.67	3.00	−5.67	−9.64	−12.82	−16.00	−8.50	−11.00	−15.29	−16.71
125 m	130	135	140	145	150	155	160	165	170	175
−11.33	−8.00	24.14	38.13	40.00	35.00	29.63	24.00	26.50	28.75	30.63
−10.33	−2.00	23.71	36.63	41.00	33.14	24.38	15.00	20.63	26.25	31.88
180 m	185	190	195	200	205	210	215	220	225	230
30.50	31.00	36.00	31.63	21.00	12.25	3.00	−7.00	−4.43	−0.50	0.75
31.00	29.13	29.75	24.75	16.00	7.25	3.25	6.38	11.57	12.25	5.38

(4) *Observed covariance functions*: To see what autocorrelation functions look like in practice, consider the data in Table 5.1. Here, we have tabulated the calibrated acoustic backscatter anomaly measured at 5-m depth increments in the upper ocean using a towed 150 kHz acoustic Doppler current profiler (ADCP). These "time series" data are from the first bin of adjacent beams 1 and 2 of a four-beam ADCP, and represent the backscatter intensity from zooplankton located at a distance of 5 m from the instrument. Since the transducers are tilted at an angle of 30° to the vertical, the two profiles are separated horizontally by only 3.5 m and the autocorrelations should be nearly identical at all lags. In this case, we use the normalized covariance (5.3.7) derived from equation (5.3.3) in which the sum is divided by the number of lag values, $N - k$, for lag $\tau = k\Delta t$.

As indicated by the autocorrelation functions in Figure 5.3.2, the functions are similar at small lags where statistical reliability is large but diverge significantly at higher lags with the decrease in the number of independent covariance estimates.

(5) *Integral time scales*: The integral time scale, T^*, is defined as the sum of the normalized autocorrelation function (5.3.7) over the length $L = N\Delta\tau$ of the time series for N lag steps, $\Delta\tau$. Specifically, the estimate

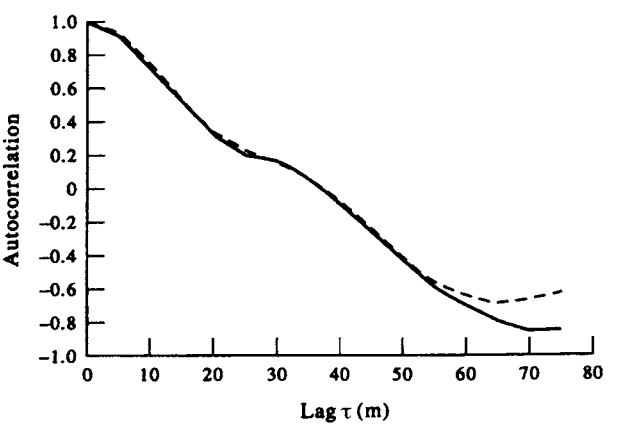


Figure 5.3.2. Autocorrelation functions of the acoustic backscatter data in Table 5.1.

$$T^* = \frac{\Delta\tau}{2}\sum_{i=0}^{N'}[\rho(\tau_i) + \rho(\tau_{i+1})]$$
$$= \frac{\Delta\tau}{2\sigma^2}\sum_{i=0}^{N'}[C(\tau_i) + C(\tau_{i+1})] \tag{5.3.16}$$

for $N' \leq N - 1$ gives a measure of the dominant correlation time scale within a data series—for times longer than T^*, the data become decorrelated. There are roughly $\Delta\tau N/T^*$ actual degrees of freedom within the time series. In reality, the summation typically is limited to $N' \ll N$ since low frequency components within the time series prevent the summation from converging to a constant value over the finite length of the record. In general, one should continue the summation until it reaches a near-constant value which we take as the value for T^*. If no plateau is reached within a reasonable number of lags, no integral time scale exists. In that case, the integral time scale can be approximated by integrating only to the first zero crossing of the autocorrelation function (cf. Poulain and Niiler, 1989).

(6) *Correlation analysis versus linear regression*: Geophysical data are typically obtained from random temporal sequences or spatial fields that cannot be regarded as mutually independent. Because the data series depend on time and/or spatial coordinates, the use of linear regression to study relationships between data series may lead to incomplete or erroneous conclusions. As an example, consider two time series: A white-noise series, consisting of identically distributed and mutually independent random variables, and the same series but with a time shift. As the values of the time series are statistically independent, the cross-correlation coefficient will be zero at zero lag, even though the time series are strictly linearly related. Regression analysis would show no relationship between the two series. However, cross-correlation analysis would reveal the linear relationship (a coefficient of unity) for a lag equal to the time shift. Correlation analysis is often a better way to study relations among time series than traditional regression analysis.

5.4 FOURIER ANALYSIS

For many applications, we can view time series as linear combinations of periodic or quasi-periodic components that are superimposed on a long-term trend and random high-frequency noise. The periodic components are assumed to have fixed, or slowly varying amplitudes and phases over the length of the record. The trends might include a slow drift in the sensor characteristics or a long-term component of variability that cannot be resolved by the data series. "Noise" includes random contributions from the instrument sensors and electronics, as well as frequency components that are outside the immediate range of interest (e.g. small-scale turbulence). A goal of time-series analysis in the frequency domain is to reliably separate periodic oscillations from the random and aperiodic fluctuations. Fourier analysis is one of the most commonly used methods for identifying periodic components in near-stationary time-series oceanographic data. (If the time series are strongly nonstationary, more localized transforms such as the Hilbert and Wavelet transforms should be used.)

The fundamentals of Fourier analysis were formalized in 1807 by the French mathematician Joseph Fourier (1768–1830) during his service as an administrator under Napoleon. Fourier developed his technique to solve the problem of heat conduction in a solid with specific application to heat dissipation in blocks of metal being turned into cannons. Fourier's basic premise was that any finite length, infinitely repeated time series, $y(t)$, defined over the principal interval $[0, T]$ can be reproduced using a linear summation of cosines and sines, or *Fourier series*, of the form

$$y(t) = \overline{y(t)} + \sum_p [A_p \cos(\omega_p t) + B_p \sin(\omega_p t)] \tag{5.4.1}$$

in which $\bar{y}$ is the mean value of the record, A_p, B_p are constants (the Fourier coefficients), and the specified angular frequencies, ω_p, are integer ($p = 1, 2, \ldots$) multiples of the fundamental frequency, $\omega_1 = 2\pi f_1 = 2\pi/T$, where T is the total length of the time series. Provided enough of these Fourier components are used, each value of the series can be accurately reconstructed over the principal interval. By the same token, the relative contribution a given component makes to the total variance of the time series is a measure of the importance of that particular frequency component in the observed signal. This concept is central to spectral analysis techniques. Specifically, the collection of Fourier coefficients having amplitudes A_p, B_p form a *periodogram* which then defines the contribution that each oscillatory component ω_p makes to the total "energy" of the observed oceanic signal. Thus, we can use the Fourier components to estimate the power spectrum (energy per unit frequency bandwidth) of a time series. Since both A_p, B_p must be specified, there are two degrees of freedom per spectral estimate derived from the "raw" or unsmoothed periodogram.

5.4.1 Mathematical formulation

Let $y(t)$ denote a continuous, finite-amplitude time series of finite duration. Examples include hourly sea-level records from a coastal tide gauge station or temperature records from a moored thermistor chain. If y is periodic, there is a period T such that $y(t) = y(t + T)$ for all t. In the language of Fourier analysis, the periodic functions are sines and cosines, which have the important properties that:

(1) A finite number of Fourier coefficients achieves the minimum mean square error between the original data and a functional fit to the data series;
(2) the functions are orthogonal so that coefficients for a given frequency can be determined independently.

Suppose that the time series is specified only at discrete times by subsampling the continuous series $y(t)$ at a sample spacing of Δt (Figure 5.4.1). Since the series has a duration T, there are a total of $N = T/\Delta t$ sample intervals and $N + 1$ sample points located at times $y(t_n) = y(n\Delta t) \equiv y_n$ ($n = 0, 1, \ldots, N$). Using Fourier analysis, it is possible to reproduce the original signal as a sum of sine or cosine waves of different amplitudes and phases. In Figure 5.4.1, we show a time series $y(n\Delta t)$ of 41 data points followed by plots of the first, second, and sixth harmonics that were summed to create the time series. The frequencies of these harmonics are $f = 1/T$, $2/T$, and $6/T$, respectively, and each harmonic has the form $y_k(n\Delta t) = C_k \cos[(2\pi kn/N + \phi_k]$ where (C_k, ϕ_k) are the amplitudes and phases of the harmonics for $k = 1, 2, 6$. Here, $T = 40\Delta t$ and we have arbitrarily chosen $(C_1, \phi_1) = (2, \pi/4)$, $(C_2, \phi_2) = (0.75, \pi/2)$,

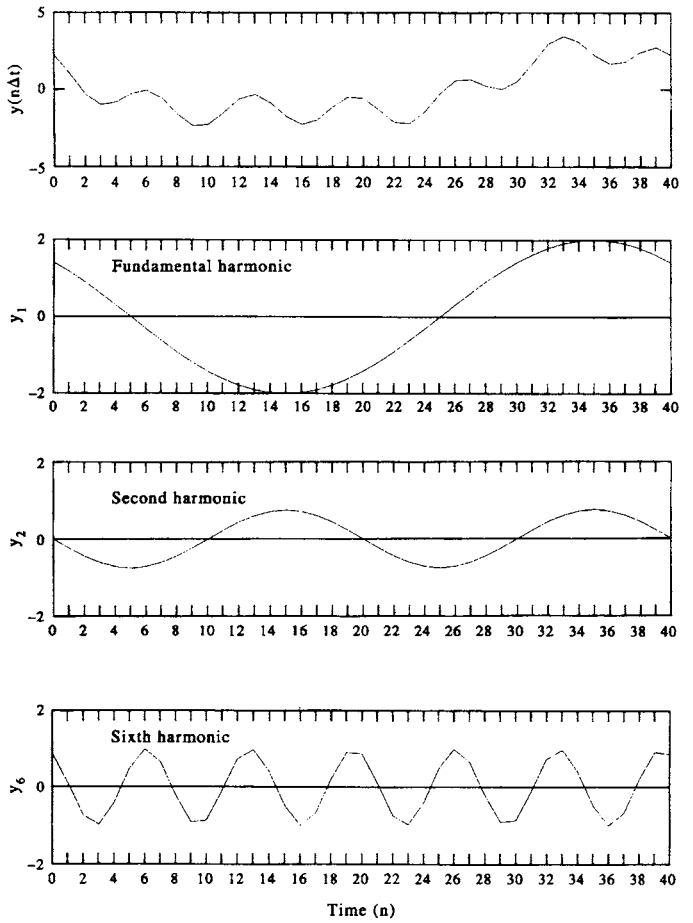


Figure 5.4.1. Discrete subsampling of a continuous signal y(t). The sampling interval is $\Delta t = 1$ time unit and the fundamental frequency is $f_1 = 1/T$ where $T = N\Delta t$ is the total record length and $N = 40$. The signal $y(t)$ is the sum of the first, second, and sixth harmonics which have the form $y_k(n\Delta t) = C_k \cos[(2\pi kn/N) + \phi_k]$; $k = 1, 2, 6$; $n = 0, 1, \ldots, 40$.

and $(C_6, \phi_6) = (1.0, \pi/6)$. The $N/2$ harmonic, which is the highest frequency component that can be resolved by this sampling, has a frequency $f_N = (N/2)/N\Delta t = 1/2\Delta t$ cycles per unit time and a period of $2\Delta t$. Called the *sampling* or *Nyquist* frequency, this represents the highest frequency resolved by the sample series in question. (In Chapter 5, we have used the subscript N to denote the Nyquist frequency and there is no confusion between this property and the integer N, as in $n = 1, 2, \ldots, N$, or the buoyancy frequency $N(z)$.) (In this chapter, we have used the subscript N to denote the Nyquist frequency f_N, and there should be no confusion between this property and the integer N, as in $n = 1, 2, \ldots, N$, or the buoyancy frequency, $N(z)$.)

The fundamental frequency, $f_1 = 1/T$, is used to construct $y(t)$ through the infinite *Fourier series*

$$y(t) = \tfrac{1}{2}A_0 + \sum_{p=1}^{\infty} [A_p \cos(\omega_p t) + B_p \sin(\omega_p t)] \tag{5.4.2}$$

in which

$$\omega_p = 2\pi f_p = 2\pi p f_1 = 2\pi p/T; \quad p = 1, 2, \ldots \tag{5.4.3}$$

is the frequency of the pth constituent in radians per unit time (f_p is the corresponding frequency in cycles per unit time) and $A_0/2$ is the mean, or "DC" offset, of the time series. The factor of 1/2 multiplying A_0 is for mathematical convenience. Note that the mean value is synonymous with the zero-frequency component obtained in the limit $\omega \to 0$. Also, the length of the data record, T, defines both the lowest frequency, f_1, resolvable by the data series and the maximum frequency resolution, $\Delta f = 1/T$, one can obtain from discretely sampled data.

To obtain the coefficients A_p, we simply multiply equation (5.4.2) by $\cos(\omega_p t)$ then integrate over all possible frequencies. The coefficients B_p, are obtained in the same way by multiplying by $\sin(\omega_p t)$. Using the orthogonality condition for the product of trigonometric functions (which requires that the trigonometric arguments cover an exact integer number of 2π cycles over the interval $(0, T)$), we find

$$A_p = \frac{2}{T}\int_0^T y(t)\cos(\omega_p t)\,dt, \quad p = 0, 1, 2, \ldots \tag{5.4.4a}$$

$$B_p = \frac{2}{T}\int_0^T y(t)\sin(\omega_p t)\,dt, \quad p = 1, 2, \ldots \tag{5.4.4b}$$

where the integral for $p = 0$ in (5.4.4a) yields $A_0 = 2\bar{y}$, twice the mean value of $y(t)$ for the entire record. Since each pair of coefficients (A_p, B_p) is associated with a frequency ω_p (or f_p), the amplitudes of the coefficients provide a measure of the relative importance of each frequency component to the overall signal variability. For example, if $(A_6^2 + B_6^2)^{1/2} \gg (A_2^2 + B_2^2)^{1/2}$ we expect there is much more "spectral energy" at frequency ω_6 than at frequency ω_2. Here, spectral energy refers to the amplitudes squared of the Fourier coefficients which represent the variance, and therefore the energy, for that portion of the time series.

We can also express our Fourier series as amplitude and phase functions in the compact Fourier series form

$$y(t) = \tfrac{1}{2}C_0 + \sum_{p=1}^{\infty} C_p \cos(\omega_p t - \theta_p) \tag{5.4.5}$$

in which the amplitude of the pth component is

$$C_p = (A_p^2 + B_p^2)^{1/2}, \quad p = 0, 1, 2, \ldots \tag{5.4.6}$$

where $C_0 = A_0$ ($B_0 = 0$) is twice the mean value and

$$\theta_p = \tan^{-1}[B_p/A_p], \quad p = 1, 2, \ldots \tag{5.4.7}$$

is the phase angle of the constituent at time $t = 0$. The phase angle gives the relative "lag" of the component in radians (or degrees) measured counterclockwise from the real axis ($B_p = 0, A_p > 0$). The corresponding time lag for the pth component is then $t_p = \theta_p/2\pi f_p$ in which θ_p is measured in radians.

The discrimination of signal amplitude as a function of frequency given by equations (5.4.2) and (5.4.5) provides us with the beginnings of spectral analysis. Notice that neither of these expressions allows for a trend in the data. If any trend is not first removed from the record, the analysis will erroneously blend the variance from the trend into the lower frequency components of the Fourier expansion. Moreover, we now see the need for the factor of 1/2 in the leading terms of (5.4.2) and (5.4.5). Without it, the $p = 0$ components would equal twice the mean component, $\bar{y} = \frac{1}{2}A_0 = \frac{1}{2}C_0$.

Up to now we have assumed that $y(t)$ is a scalar quantity. We can also expand the time series of a vector property, $\mathbf{u}(t)$. Included in this category are time series of current velocity from moored current meter arrays and wind velocity from moored weather buoys. Expressing vector time series in complex notation, we can write

$$\mathbf{u}(t) = u(t) + \mathrm{i}v(t) \tag{5.4.8}$$

where, for example, u and v might be the north–south and east–west components of current velocity in Cartesian coordinates. An individual vector can be expressed as

$$\mathbf{u}(t) = \overline{\mathbf{u}(t)} + \sum_{p=1}^{\infty} [A_p \cos(\omega_p t + \alpha_p) + \mathrm{i}B_p \sin(\omega_p t + \beta_p)] \tag{5.4.9}$$

Here, $\overline{\mathbf{u}(t)}$ is the mean (time averaged) vector, $\bar{\mathbf{u}} = \bar{u} + \mathrm{i}\bar{v}$, and (α_p, β_p) are phase lags or relative phase differences for the separate velocity components.

Vector quantities also can be defined through expressions of the form

$$\begin{aligned}\mathbf{u}(t) = \bar{\mathbf{u}} + \sum_{p=1}^{\infty} \{ & \exp[\mathrm{i}(\varepsilon_p^+ + \varepsilon_p^-)/2][(A_p^+ + A_p^-)\cos[\omega_p t + (\varepsilon_p^+ - \varepsilon_p^-)/2] \\ & + \mathrm{i}(A_p^+ - A_p^-)\sin[\omega_p t + (\varepsilon_p^+ - \varepsilon_p^-)/2]\}\end{aligned} \tag{5.4.10}$$

in which A_p^+ and A_p^- are, respectively, the lengths of the counterclockwise (+) and clockwise (−) rotary components of the velocity vector, and ε_p^+ and ε_p^- are the angles that these vectors make with the real axis at $t = 0$. The resultant time series is an ellipse with major axis of length $L_M = A_p^+ + A_p^-$ and minor axis of length $L_m = |A_p^+ - A_p^-|$. The major axis is oriented at angle $\theta_p = \frac{1}{2}(\varepsilon_p^+ + \varepsilon_p^-)$ from the u-axis and the current rotates counterclockwise when $A_p^+ > A_p^-$ and clockwise when $A_p^+ < A_p^-$. The velocity vector is aligned with the major axis direction θ_p when $\omega_p t = -\frac{1}{2}(\varepsilon_p^+ - \varepsilon_p^-)$. Motions are said to be *linearly polarized* (rectilinear) if the two oppositely rotating components are of the same magnitude and *circularly polarized* if one of the two components is zero. In the northern (southern) hemisphere, motions are predominantly clockwise (counterclockwise) rotary. Further details on rotary decomposition are presented in Sections 5.6 and 5.8.

5.4.2 Discrete time series

Most oceanographic time or space series, whether they were collected in analog or digital form, are eventually converted to digital data which may then be expressed as series expansions of the form (5.4.2) or (5.4.5). These expansions are then used to compute the Fourier transform (or periodogram) of the data series. The basis for this transform is Parseval's theorem which states that the mean square (or average) energy

of a time series $y(t)$ can be separated into contributions from individual harmonic components to make up the time series. For example, if $\bar{y}$ is the sample mean value of the time series, y_p is the contribution from the nth data value and N is the total number of data values in the time series, then the mean square value of the series about its mean (i.e. the variance of the time series)

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}[y_n - \bar{y}]^2 \tag{5.4.11}$$

provides a measure of the total energy in the time series. The variance (5.4.11) also can be obtained by summing the contributions from the individual Fourier harmonics. This kind of decomposition of discrete time series into specific harmonics leads to the concept of a Fourier line spectrum (Figure 5.4.2).

To determine the energy distribution within a time series, $y(t)$, we need to find its Fourier transform. That is, we need to determine the coefficients A_p, B_p in the Fourier series (5.4.2) or, equivalently, the amplitudes and phase lags, C_p, θ_p in the Fourier series (5.4.5). Suppose that we have first removed any trend from the data record. For any time t_n, the Fourier series for a finite length, de-trended digital record having N (even) values at times $t_n = t_1, t_2, \ldots, t_N$, is

$$y(t_n) = \tfrac{1}{2}A_0 + \sum_{p=1}^{N/2}[A_p \cos(\omega_p t_n) + B_p \sin(\omega_p t_n)] \tag{5.4.12}$$

where the angular frequency $\omega_p = 2\pi f_p = 2\pi p/T$. Using $t_n = n \cdot \Delta t$ together with (5.4.6) and (5.4.7), the final form for the discrete, finite Fourier series becomes

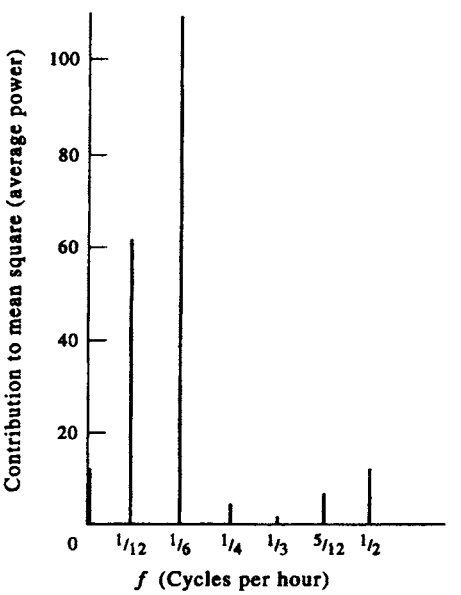


Figure 5.4.2. An example of a Fourier line spectrum with power at discrete frequencies, f, for a 24-h duration record with 1-h sampling increment.

$$y(t_n) = \tfrac{1}{2}A_0 + \sum_{p=1}^{N/2}\left[A_p \cos(2\pi pn/N) + B_p \sin(2\pi pn/N)\right]$$
$$= \tfrac{1}{2}C_0 + \sum_{p=1}^{N/2} C_p \cos\left[(2\pi pn/N) - \theta_p\right] \tag{5.4.13}$$

where the leading terms, $\frac{1}{2}A_0$ and $\frac{1}{2}C_0$, are the mean values of the record. The coefficients are again determined using the orthogonality condition for the trigonometric functions. In fact, the main difference between the discrete case and the continuous case formulated in the last section (aside from the fact we can no longer have an infinite number of Fourier components) is that coefficients are now defined through the summations rather than through integrals

$$A_p = \frac{2}{N}\sum_{n=1}^{N} y_n \cos(2\pi pn/N), \quad p = 0, 1, 2, \ldots, N/2$$
$$A_0 = \frac{2}{N}\sum_{n=1}^{N} y_n, \quad B_0 = 0$$
$$A_{N/2} = \frac{1}{N}\sum_{n=1}^{N} y_n \cos(n\pi), \quad B_{N/2} = 0$$
$$B_p = \frac{2}{N}\sum_{n=1}^{N} y_n \sin(2\pi pn/N), \quad p = 1, 2, \ldots, (N/2) - 1 \tag{5.4.14}$$

Notice that the summations in equations (5.4.14) consist of multiplying the data record by sine and cosine functions which "pick out" from the record those frequency components specific to their trigonometric arguments. Remember, the orthogonality condition requires that the arguments in the trigonometric functions be integer multiples of the total record length, $T = N\Delta t$, as they are in equation (5.4.14). If they are not, the sines and cosines do not form an orthonormal set of basis functions for the Fourier expansion and the original signal cannot be correctly replicated.

The arguments $2\pi pn/N$ in the above equations are based on a hierarchy of equally spaced frequencies $\omega_p = 2\pi p/(N\Delta t)$ and time increment "n". The summation goes to $N/2$ which is the limit of coefficients we can determine; for $p > N/2$ the trigonometric functions simply begin to cause repetition of coefficients already obtained for the interval $p \leq N/2$. Furthermore, it should be obvious that because there are as many coefficients as data points and because the trigonometric functions form an orthogonal basis set, the summation over the $2(N/2) = N$ discrete coefficients provides an exact replication of the time series, $y(t)$. Small differences between the original data and the Fourier series representation arise because of roundoff errors accumulated during the arithmetic calculations (see Chapter 3).

The steps in computing the Fourier coefficients are as follows. Step 1: calculate the arguments $\Phi_{pn} = 2\pi pn/N$ for each integer p and n. Step 2: for each $n = 1, 2, \ldots, N$, evaluate the corresponding values of $\cos\Phi_{pn}$ and $\sin\Phi_{pn}$, and collect sums of $y_n \cdot \cos\Phi_{pn}$ and $y_n \cdot \sin\Phi_{pn}$. Step 3: Increment p and repeat steps 1 and 2. The procedure requires roughly N^2 real multiply-add operations. For any real data sequence, roundoff errors plus errors associated with truncation of the total allowable number of

desired Fourier components (maximum $f_p < f_{N/2}$) will give rise to a less than perfect fit to the data. The residual $\Delta y(t) = y(t) - y_{FS}(t)$ between the observations $y(t)$ and the calculated Fourier series $y_{FS}(t)$ will diminish with increased computational precision and increased numbers of allowable terms used in the series expansion. When computing the phases $\theta_p = \tan^{-1}[B_p/A_p]$ in the formulation (5.4.13), one must take care to examine in which quadrants A_p and B_p are situated. For example, $\tan^{-1}(0.2/0.7)$ differs from $\tan^{-1}(-0.2/-0.7)$ by 180°. The familiar ATAN2 function in FORTRAN is especially designed to take care of this problem.

5.4.3 A computational example

The best way to demonstrate the computational procedure for Fourier analysis is with an example. Consider the two-year segment of monthly mean sea surface temperatures measured at the Amphitrite light station off the southwest coast of Vancouver Island (Table 5.2). Each monthly value is calculated from the average of daily surface thermometer observations collected around noon local time and tabulated to the nearest 0.1°C. These data are known to contain a strong seasonal cycle of warming and cooling which is modified by local effects of runoff, tidal stirring and wind mixing.

The data in Table 5.2 are in the form $y(t_n)$, where $n = 1, 2, \ldots, N$ ($N = 24$). To calculate the coefficients A_p and B_p for these data, we use the summations (5.4.14) for each successive integer p, up to $p = N/2$. These coefficients are then used in (5.4.6) to calculate the magnitude $C_p = (A_p^2 + B_p^2)^{1/2}$ for each frequency component, $f_p = p/T$. Since C_p^2 is proportional to the variance at the specified frequency, the C_p enable us to rate the order of importance of each frequency component in the data series.

The mean value $\overline{y(t)} = \frac{1}{2}A_0$ and the 12 pairs of Fourier coefficients obtainable from the temperature record are listed in Table 5.3 together with the magnitude C_p. Values have been rounded to the nearest 0.01°C. The Nyquist frequency, f_N, is 0.50 cycles per month (cpmo, $p = 12$) and the fundamental frequency, f_1, is 0.042 cpmo ($p = 1$). As we would anticipate from a visual inspection of the time series, the record is dominated by the annual cycle (period = 12 months) followed by weaker contributions from the bi-annual cycle (24 months) and semi-annual cycle (six months). For periods shorter than six months, the coefficients C_p have similar amplitudes and likely represent the roundoff errors and background "noise" in the data series. This suggests that we can reconstruct the original time series to a high degree of accuracy using only the mean value ($p = 0$) and the first three Fourier coefficients ($p = 1, 2, 3$).

Figure 5.4.3 is a plot of the original sea surface temperature (SST) time series and the reconstructed Fourier fit to this series using only the first three Fourier components from Table 5.3. Comparison of these two time series, shows that the reconstructed series does not adequately reproduce the skewed crest of the first year nor the high-frequency "ripples" in the second year of the data record. There also is a

Table 5.2. Monthly mean sea surface temperatures SST (°C) at Amphitrite Point (48°55.16'N, 125°32.17'W) on the west coast of Canada for January 1982 through December 1983

Year 1982												
n	1	2	3	4	5	6	7	8	9	10	11	12
SST	7.6	7.4	8.2	9.2	10.2	11.5	12.4	13.4	13.7	11.8	10.1	9.0
Year 1983												
n	13	14	15	16	17	18	19	20	21	22	23	24
SST	8.9	9.5	10.6	11.4	12.9	12.7	13.9	14.2	13.5	11.4	10.9	8.1

Table 5.3. Fourier coefficients and frequencies for the Amphitrite Point monthly mean temperature data. Frequency is in cycles per month (cpmo). $A_0/2$ is the mean temperature and θ_p is the phase lag for the pth component taken counterclockwise from the positive A_p axis

p	Freq. (cpmo)	Period (month)	Coeff. A_p (°C)	Coeff. B_p (°C)	Coeff. C_p (°C)	Phase θ_p (degrees)
0	0	–	21.89	0	21.89	0
1	0.042	24	−0.55	−0.90	1.05	−121.4
2	0.083	12	−1.77	−1.99	2.67	−131.7
3	0.125	8	0.22	−0.04	0.23	−10.3
4	0.167	6	−0.44	−0.06	0.45	−172.2
5	0.208	4.8	0.09	−0.07	0.11	−37.9
6	0.250	4	0.08	−0.04	0.09	−26.6
7	0.292	3.4	0.01	−0.16	0.16	−58.0
8	0.333	3	−0.03	−0.16	0.16	−100.6
9	0.375	2.7	−0.14	0.05	0.15	160.3
10	0.417	2.4	−0.09	−0.07	0.11	−142.1
11	0.458	2.2	−0.08	−0.12	0.14	−123.7
12	0.500	2	−0.15	0	0.15	0

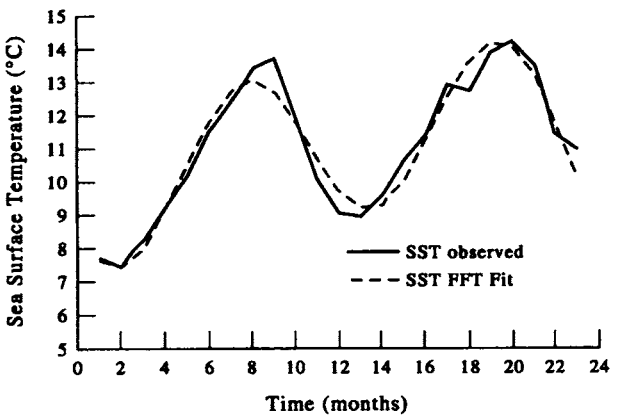


Figure 5.4.3. Monthly mean sea surface temperature (SST) record for Amphitrite Point on the west coast of Vancouver Island (see Table 5.1). The bold line is the original 24-month series; the dashed line is the SST time series generated using the first three Fourier components, f_p, $p = 0, 1, 2$, corresponding to the mean, 24-month, and 12-month cycles (Fourier components appear in Table 5.2).

slight mismatch in the maxima and minima between the series. Differences between the two curves are typically around a few tenths of a degree. In contrast, if we use all 12 components in Table 5.3, corresponding to 24 degrees of freedom, we get an exact replica of the original time series to within machine accuracy.

5.4.4 Fourier analysis for specified frequencies

Analysis of time series for specific frequencies is a special case of Fourier analysis that involves adjustment of the record length to match the periods of the desired Fourier components. As we illustrate in the following sections, analysis for specific frequency components is best conducted using least-squares fitting methods rather than Fourier analysis. Least-squares analysis requires that there be many fewer constituents than data values, which is usually the case for tidal analysis at the well-defined frequencies

of the tide-generating potential. Problems arise if there are too few data values. For example, suppose that we have a few days of hourly water level measurements and we want to use Fourier analysis to determine the amplitudes and phases of the daily tidal constituents, f_k. To do this, we need to satisfy the orthogonality condition for the trigonometric basis functions for which terms like $\int \cos(2\pi f_j t)\cos(2\pi f_k t)\,dt$ are zero except where $f_j = f_k$ (the integral is over the entire length of the record, T). The approach is only acceptable when the length of the data set is an integer multiple of all the harmonic frequencies we are seeking. That is, the specified tidal frequencies f_k must be integer multiples of the fundamental frequency, $f_1 = 1/T$, such that $f_k \cdot T = 1, 2, \ldots,$ If this holds, we can use Fourier analysis to find the constituent amplitudes and phases at the specified frequencies. In fact, this integer constraint on $f_k \cdot T$ is a principal reason why oceanographers prefer to use record lengths of 14, 29, 180, or 355 days when performing analyses of tides. Since the periods of most of the major tidal constituents (K_1, M_2, etc.) are integer multiples of the fundamental tidal periods (one lunar day, one lunar month $\approx$ 29 days, one year, 8.8 years, 18.6 years, etc.) of the above record lengths, the analysis is aided by the orthogonality of the trigonometric functions.

A note for those unfamiliar with tidal analysis terminology: Letters of tidal harmonics identify the different types ("species") of tide in each frequency band. Harmonic components of the tide-producing force that undergo one cycle per lunar day ($\approx$ 25 h) have a subscript 1 (e.g. K_1), those with two cycles per lunar day have subscript 2 (e.g. M_2), and so on. Constituents having one cycle per day are called diurnal constituents, those with two cycles per day, semidiurnal constituents. The main daily tidal component, the K_1 constituent, has a frequency of 0.0418 cph (corresponding to an angular speed of 15.041° per mean solar hour) and is associated with the cyclic changes in the luni-solar declination. The main semidiurnal tidal constituent, the M_2 constituent, has a frequency of 0.0805 cph (corresponding to an angular speed of 28.984° per mean solar hour) and is associated with cyclic changes in the lunar position relative to the earth. Other major daily constituents are the O_1, P_1, S_2, N_2, and K_2 constituents. In terms of the tidal potential, the hierarchy of tidal constituents is $M_2, K_1, S_2, O_1, P_1, N_2, K_2, \ldots$. Other important tidal harmonics are the lunar fortnightly constituent, M_f, the lunar monthly constituent, M_m, and the solar annual constituent, S_a. For further details the reader is referred to Thomson (1981) and Foreman (1977).

Returning to our discussion concerning Fourier analysis at specified frequencies, consider the 32-h tide gauge record for Tofino, British Columbia presented in Figure 5.4.4. As we show in Section 5.5, least-squares analysis can be used to reproduce this short record quite accurately using only the K_1 tidal constituent and the M_2 constituent. These are the dominant tidal constituents in all regions of the ocean except near amphidromic points. Because the record is 32 h long, the diurnal and semidiurnal frequencies are not integer multiples of the fundamental frequency $f_1 = 1/T = 0.031$ cph and are not among the sequence of 16 possible frequencies generated from the Fourier analysis. In order to have frequency components centered more exactly at the K_1 and M_2 frequencies, we would need to shorten the record to 24 h or pad the existing record to 48 h using zeros. In either case, the $f_k \cdot T$ for the tides would then be close to integers and a standard Fourier analysis would give an accurate fit to the observed time series. If we stick with the 32-h series, we find that the tidal energy in the diurnal and semidiurnal bands is partitioned among the first three Fourier components at frequencies $f_1 = 0.031$, $f_2 = 0.062$, and $f_3 = 0.093$ cph. These

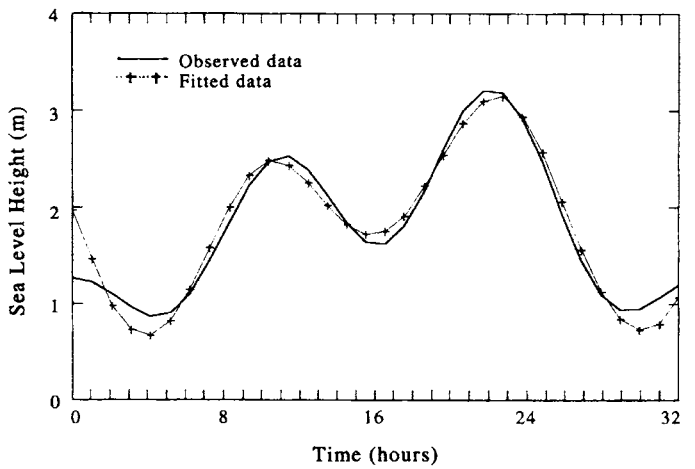


Figure 5.4.4. Hourly sea-level height (SLH) recorded at Tofino on the west coast of Vancouver Island (see Table 5.7). The bold line is the original 32-h series; the dotted line is the SLH series generated using the mean ($p = 0$) plus the next three Fourier components, f_p, $p = 1, 2, 3$ having nontidal periods T_p of 32, 16, and 8 h, respectively.

frequencies are only vaguely close to those of the diurnal and semidiurnal constituents but do span the energy-containing frequency bands. As a result, the time series generated from the record mean combined with the first three Fourier components ($p = 1, 2, 3$) closely approximates the time series obtained using the true tidal frequencies (see Figure 5.5.2).

5.4.5 The fast Fourier transform

One of the main problems with both the autocovariance and the direct Fourier methods of spectral estimation is low computational speed. The Fourier method requires the expansion into series of sine and cosine terms—a time-consuming procedure. The fast Fourier transform (FFT) is a way to speed up this computation while retaining the accuracy of the direct Fourier method. This makes the Fourier method computationally more attractive than the autocovariance approach.

To illustrate the improved efficiency of the FFT method, consider a series of N values for which $N = 2^p$ (p is a positive integer). The discrete Fourier transform of this series would require N^2 operations whereas the FFT method requires only $8N\log_2 N$ operations. The savings in computer time can be substantial. For example, if $N = 8192$, $N^2 = 67{,}108{,}864$ while $8N\log_2 N = 851{,}968$. Computers are much faster now than when the FFT method was introduced but the relative savings in computational efficiency remains the same. Bendat and Piersol (1986) define the speed ratio between the FFT and discrete Fourier method as $N/4p$. This becomes increasingly more important as the number of terms increases since the direct method computational time is $O(N^2)$ while for the FFT method it is $O(N)$. If one is seeking a smoothed power spectrum, it is often more efficient to compute the spectrum using the FFT technique and then smooth in spectral space by averaging over adjoining frequency bands rather than smoothing with an autocovariance lag window in the time domain.

To understand the FFT algorithm, we follow the derivation of Danielson and Lanczos (1942) who first helped pioneer the method. Consider a time series of x_t, where $t = 1, 2, \ldots, N$. We want to find the Fourier transform $X_m = X(m/N\Delta t)$, where $m = 0, 1, \ldots, N-1$. To do this, we first partition x_t into two half-series y_t and z_t, where $y_t = x_{2t-1}, z_t = x_{2t}, t = 1, 2, \ldots, N/2$. The series y_t contains values at the odd number times $(x_1, x_3, \ldots)$ while the function z_t contains values at the even number times $(x_2, x_4, \ldots)$. Both functions have $N/2$ values and their Fourier transforms are

$$Y_m^{(N/2)} = \frac{2}{N}\sum_{t=1}^{N/2} y_t \exp\left[\frac{(-\mathrm{i}4\pi tm)}{N}\right] \tag{5.4.15a}$$

$$Z_m^{(N/2)} = \frac{2}{N}\sum_{t=1}^{N/2} z_t \exp\left[\frac{(-\mathrm{i}4\pi tm)}{N}\right] \tag{5.4.15b}$$

where the superscript is used to denote the number of terms used in the expansion. But $X_m^{(N)}$, $Y_m^{(N/2)}$, and $Z_m^{(N/2)}$ are related since

$$\begin{aligned} X_m^{(N)} &= \frac{2}{N}\sum_{t=1}^{N/2} x_t \exp\left[\frac{-\mathrm{i}4\pi tm}{N}\right] \\ &= \frac{1}{N}\sum_{t=1}^{N/2}\left\{y_t \exp\left[\frac{-\mathrm{i}4\pi tm}{N}(2t-1)\right] + z_t \exp\left[\frac{-\mathrm{i}4\pi tm}{N}(2t)\right]\right\} \\ &= \tfrac{1}{2}\exp\left[\frac{(\mathrm{i}2\pi m)}{N}\right] Y_m^{(N/2)} + \tfrac{1}{2}Z_m^{(N/2)}, \quad 0 \le m \le (N/2)-1 \end{aligned} \tag{5.4.16}$$

Also

$$\begin{aligned} Y_{m+N/2}^{(N/2)} &= Y_m^{(N/2)}; \quad 0 \le m \le N/2-1 \\ Z_{m+N/2}^{(N/2)} &= Z_m^{(N/2)}; \quad 0 \le m \le N/2-1 \end{aligned} \tag{5.4.17}$$

so that

$$\begin{aligned} X_{m+N/2}^{(N)} &= \frac{1}{2}\exp\left[\mathrm{i}\left(\frac{2\pi}{N}\right)\left(m+\frac{N}{2}\right)\right] Y_m^{(N/2)} + \frac{1}{2}Z_m^{(N/2)} \\ &= -\frac{1}{2}\exp\left(\frac{\mathrm{i}2\pi m}{N}\right) Y_m^{(N/2)} + \frac{1}{2}Z_m^{(N/2)}, \quad 0 \le m \le (N/2)-1 \end{aligned} \tag{5.4.18}$$

thus

$$X_m^{(N)} = \frac{1}{2}\exp\left[\mathrm{i}\frac{(2\pi m)}{N}\right] Y_m^{(N/2)} + \frac{1}{2}Z_m^{(N/2)}, \quad 0 \le m \le N/2-1 \tag{5.4.19}$$

and

$$X_{m-N/2}^{(N)} = -\frac{1}{2}\exp\left[\mathrm{i}\frac{(2\pi m)}{N}\right] Y_m^{(N/2)} + \frac{1}{2}Z_m^{(N/2)}, \quad 0 \le m \le N/2-1 \tag{5.4.20}$$

Thus, the Fourier transform for the series x_t is found from the Fourier series of the half series y_t, z_t. Since $N/2$ is even, this can be repeated. If the length of the data is not

a power of 2, it should be padded with zeros up to the next power of two. For a series of length $N = 2p$ (p a positive integer), the procedure is followed until partitions consist of only one term whose Fourier transform equals itself, or the procedure is followed until N becomes a prime number, i.e. $N = 3$. The Fourier transform is then found directly for the remaining short series.

5.5 HARMONIC ANALYSIS

Standard Fourier analysis involves the computation of Fourier amplitudes at equally spaced frequency intervals determined as integer multiples of the fundamental frequency, f_1. That is, for frequencies f_1, $2f_1$, $3f_1$, ... , f_N (f_N = Nyquist frequency). However, as we have shown in the previous section, standard Fourier analysis is not much use when it comes to the analysis of data series in terms of predetermined frequencies. In the case of tidal motions, for example, it would be silly to use any frequencies except those of the astronomical tidal forces. Equally important, we want to determine the amplitudes and phases of as many frequency components as possible by using as short a time series as possible. Since there are typically many more data values than there are prescribed frequencies, we have to deal with an overdetermined problem. This leads to a form of signal demodulation known as *harmonic analysis* in which the user specifies the frequencies to be examined and applies least-squares techniques to solve for the constituents. Harmonic analysis was originally designed for the analysis of tidal variability but applies equally to analysis at the annual and semi-annual periods or any other well-defined cyclic oscillation. The familiar hierarchy of "harmonic" tidal constituents is dominated by diurnal and semidiurnal motions, followed by fortnightly, monthly, semi-annual, and annual variability. In this section, we present a general discussion of harmonic analysis. The important subject of harmonic analysis of tides and tidal currents is treated separately in Section 5.5.3.

The harmonic analysis approach yields the required amplitudes and phase lags of the harmonic tidal coefficients or any other constituents we may wish to specify. Once these coefficients have been determined, we can use them to reconstruct the original time series. In the case of tidal motions, subtraction of the reconstructed tidal signal from the original record yields a time series of the nontidal or *residual* component of the time series. In many cases, it is the residual or "de-tided" signal that is of primary interest. If we break the original time series into adjoining or overlapping segments, we can apply harmonic analysis to the segments to obtain a sequence of estimates for the amplitudes and phase lags of the various frequencies of interest. This leads to the notion of signal *demodulation*.

5.5.1 A least-squares method

Suppose we wish to determine the harmonic constituents A_q and B_q for M specified frequencies which, in general, will differ from the Fourier frequencies defined by (5.4.3). In this case, $q = 0, 1, \ldots, M$ and $B_0 = 0$ so that there are a total of $2M + 1$ harmonic coefficients. Assume that there are many more observations, N, than specified coefficients (i.e. that $2M + 1 \ll N$). The problem of fitting M harmonic curves to the digital time series is then overdetermined and must be solved using an optimization technique. Specifically, we estimate the amplitudes and phases of the various components by minimizing the squared difference (i.e. the least squares)

between the original data series and our fit to that series. The coefficients for each of the M resolvable constituents are found through solution of a $(M + 1) \times (M + 1)$ matrix equation.

For M possible harmonic constituents, the time series $x(t_n)$, $n = 1, \ldots, N$ can be expanded as

$$x(t_n) = \bar{x} + \sum_{q=1}^{M} C_q \cos(2\pi f_q t_n - \phi_q) + x_r(t_n) \tag{5.5.1}$$

in which $\overline{x(t)}$ is the mean value of the record, x_r is the residual portion of the time series (which may contain other kinds of harmonic constituents), $t_n = n\Delta t$, and where C_q, f_q and ϕ_q are respectively the constant amplitude, frequency and phase of the qth constituent that we have specified. In the present configuration, we assume that the specified frequencies have the form $f_q = q/N\Delta t$ so that the argument $2\pi f_q t_n = 2\pi qn/N$. Reformulation of equation (5.5.1) as

$$x(t_n) = \bar{x} + \sum_{q=1}^{M} [A_q \cos(2\pi f_q t_n) + B_q \sin(2\pi f_q t_n)] + x_r(t_n) \tag{5.5.2}$$

yields a representation in terms of the unknown coefficients A_q, B_q where

$$\begin{aligned} C_q &= (A_q^2 + B_q^2)^{1/2}, \quad \text{(frequency component amplitude)} \\ \phi_q &= \tan^{-1}(B_q/A_q), \quad \text{(frequency component phase lag)} \end{aligned} \tag{5.5.3}$$

for $q = 0, \ldots, M$. To reduce roundoff errors (Section 3.17.3), the mean value, $\bar{x}$, should be subtracted from the record prior to the computation of the Fourier coefficients.

The objective of the least-squares analysis is to minimize the variance, e^2, of the residual time series $x_r(t_n)$ in equation (5.5.2), where

$$e^2 = \sum_{n=1}^{N} x_r^2(t_n) = \sum_{n=1}^{N} \left\{ x(t_n) - \left[\bar{x} + \sum_{q=1}^{M} M(t_n) \right] \right\}^2 \tag{5.5.4}$$

and where for convenience we define $\sum M$ as

$$\begin{aligned} \sum_{q=1}^{M} M(t_n) &= \sum_{q=1}^{M} [A_q \cos(2\pi f_q t_n) + B_q \sin(2\pi f_q t_n)] \\ &= \sum_{q=1}^{M} [A_q \cos(2\pi qn/N) + B_q \sin(2\pi qn/N)] \end{aligned} \tag{5.5.5}$$

Taking the partial derivatives of (5.5.4) with respect to the unknown coefficients A_q and B_q, and setting the results to zero, yields $2M + 1$ simultaneous equations for the $M + 1$ constituents

$$\begin{aligned} \frac{\partial e^2}{\partial A_q} = 0 &= 2\sum_{n=1}^{N} \left\{ \left[x_n - \left(\bar{x} + \sum M \right) \right] [-\cos(2\pi qn/N)] \right\}, \quad k = 0, \ldots, M \\ \frac{\partial e^2}{\partial B_q} = 0 &= 2\sum_{n=1}^{N} \left\{ \left[x_n - \left(\bar{x} + \sum M \right) \right] [-\sin(2\pi qn/N)] \right\}, \quad k = 1, \ldots, M \end{aligned} \tag{5.5.6}$$

Derivation of the coefficients in (5.5.6) requires solution of a matrix equation of the form $\mathbf{Dz} = \mathbf{y}$ in which $\mathbf{D}$ is an $(M + 1) \times (M + 1)$ matrix involving sine and cosine summation terms, $\mathbf{y}$ is a vector (column matrix) incorporating summations over the data series and $\mathbf{z}$ is a column matrix containing the required coefficients A_q and B_q. Gaps in the data are still permitted at this stage since the observation times, t_n, used in the least-squares method are not required to be evenly spaced.

Details on the matrix inversion and related problems can be found in Foreman (1977). To simplify the summations (5.5.6), trigonometric identities are often used. This requires that the data be evenly spaced and that the matrix terms be calculated over segments of the time series with no gaps. The resultant matrix $\mathbf{D}$ is symmetric so that only the upper triangle consisting of $2M + 3M + 1$ elements needs to be stored during the computations. We then seek solutions z through the matrix equation

$$\mathbf{z} = \mathbf{D}^{-1}\mathbf{y} \tag{5.5.7}$$

where $\mathbf{D}^{-1}$ is the inverse of the matrix

$$\mathbf{D} = \begin{pmatrix} N & c_1 & c_2 & \cdots & c_M & s_1 & s_2 & \cdots & s_M \\ c_1 & cc_{11} & cc_{12} & \cdots & cc_{11M} & cs_{11} & cs_{12} & \cdots & cs_{1M} \\ c_2 & cc_{21} & cc_{22} & \cdots & cc_{2M} & cs_{21} & cs_{22} & \cdots & cs_{2M} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_M & cc_{M1} & cc_{M2} & \cdots & cc_{MM} & cs_{M1} & cs_{M2} & \cdots & cs_{MM} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_1 & sc_{11} & sc_{12} & \cdots & sc_{1M} & ss_{11} & ss_{12} & \cdots & ss_{1M} \\ s_2 & sc_{21} & sc_{22} & \cdots & sc_{2M} & ss_{21} & ss_{22} & \cdots & ss_{2M} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_M & sc_{M1} & sc_{M2} & \cdots & sc_{MM} & ss_{M1} & ss_{M2} & \cdots & ss_{MM} \end{pmatrix} \tag{5.5.8}$$

and $\mathbf{y}$ and $\mathbf{z}$ are column vectors

$$\mathbf{y} = \begin{pmatrix} yc_o \\ yc_1 \\ yc_2 \\ \cdots \\ \cdots \\ yc_M \\ ys_1 \\ \cdots \\ ys_M \end{pmatrix} \quad \text{and} \quad \mathbf{z} = \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ \cdots \\ \cdots \\ A_M \\ B_1 \\ \cdots \\ B_M \end{pmatrix} \tag{5.5.9}$$

The elements of $\mathbf{z}$ yield the required coefficients A_q, B_q for each specified harmonic constituent. To find these solutions, we substitute the elements of $\mathbf{D}$ for times $t_n = n\Delta t$ and, using $\alpha_k = f_k T$, $\alpha_j = f_j T$, where f_k and f_j are frequency units of Δt^{-1} and $T = N\Delta t$ is the record length.

$$
\begin{aligned}
c_k &= \sum_{n=1}^{N} \cos(2\pi\alpha_k n/N), \quad s_k = \sum_{n=1}^{N} \sin(2\pi\alpha_k n/N) \\
cc_{kj} &= cc_{jk} = \sum_{n=1}^{N} \left[\cos(2\pi\alpha_k n/N)\cos(2\pi\alpha_j n/N)\right] \\
ss_{kj} &= ss_{jk} = \sum_{n=1}^{N} \left[\sin(2\pi\alpha_k n/N)\sin(2\pi\alpha_j n/N)\right] \\
cs_{kj} &= sc_{jk} = \sum_{n=1}^{N} \left[\cos(2\pi\alpha_k n/N)\sin(2\pi\alpha_j n/N)\right]
\end{aligned}
\tag{5.5.10}
$$

where $\alpha_k n/N = (\alpha_k/N\Delta t)(n\Delta t)$, and the elements of $\mathbf{y}$ are given by

$$
yc_k = \sum_{n=1}^{N} x_n \cos(2\pi\alpha_k n/N), \quad ys_k = \sum_{n=1}^{N} x_n \sin(2\pi\alpha_k n/N) \tag{5.5.11}
$$

5.5.2 A computational example

We can illustrate the power of the least-squares method by again using the monthly mean sea surface temperature record of Table 5.2. Our purpose is to estimate the amplitudes and phases of the dominant annual and semi-annual constituents in the Tofino temperature record and compare the results with those we obtained using Fourier analysis in Section 5.4.3. This is also the approach we would use if we wanted to subtract these particular components from the original data record, as we might want to do prior to consideration of less dominant higher frequency variability or before cross-correlation with another data set. We let $f_1 = 1/12$ month (= 0.0833 cpmo) and $f_2 = 1/6$ month (= 0.1667 cpmo) represent the frequencies of interest. From (5.5.8) and (5.5.10), we find for $\alpha_1 = f_1 T = \frac{1}{12} \times 24 = 2$, and $\alpha_2 = f_2 T = \frac{1}{6} \times 24 = 4$ that

$$
\mathbf{D} = \begin{pmatrix}
N & c_1 & c_2 & s_1 & s_2 \\
c_1 & cc_{11} & cc_{12} & cs_{11} & cs_{12} \\
c_2 & cc_{21} & cc_{22} & cs_{21} & cs_{22} \\
s_1 & sc_{11} & sc_{12} & ss_{11} & ss_{12} \\
s_2 & sc_{21} & sc_{22} & ss_{21} & ss_{22}
\end{pmatrix} \tag{5.5.12}
$$

$$
= \begin{pmatrix}
24 & 0 & 0 & 0 & 0 \\
0 & 12 & 0 & 0 & 0 \\
0 & 0 & 12 & 0 & 0 \\
0 & 0 & 0 & 12 & 0 \\
0 & 0 & 0 & 0 & 12
\end{pmatrix} \tag{5.5.13}
$$

and from (5.5.9) and (5.5.11)

$$\mathbf{y} = \begin{pmatrix} yc_0 \\ yc_1 \\ yc_2 \\ ys_1 \\ ys_2 \end{pmatrix} = \begin{pmatrix} 262.70 \\ -21.30 \\ -5.30 \\ -23.87 \\ -0.69 \end{pmatrix} \tag{5.5.14}$$

where the elements of $\mathbf{y}$ have units of °C. The solution $\mathbf{z} = \mathbf{D}^{-1}\mathbf{y}$ is the vector

$$\mathbf{z} = \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} 10.95 \\ -1.77 \\ -0.44 \\ -1.99 \\ -0.06 \end{pmatrix} \tag{5.5.15}$$

with units of °C. The results are summarized in Table 5.4. As required, the amplitudes and phases of the annual and semi-annual constituents are identical to those obtained using Fourier analysis (see Table 5.3). A plot of the original temperature record and the least-squares fitted curve using the annual and semi-annual constituents is presented in Figure 5.5.1. The standard deviation for the original record is 2.08°C

Table 5.4. Coefficients for the annual and semi-annual frequencies from a least-squares analysis of the Amphitrite Point monthly mean temperature series (Table 5.2). Frequency units are cycles per month (cpmo). q = 0 gives the mean value for the 24-month record. Other coefficients are defined through equation (5.5.3)

q	Frequency (cpmo)	Period (month)	A_q (°C)	B_q (°C)	C_q (°C)
0	–	–	10.95	0.0	10.95
2	0.083	12	−1.77	−1.99	2.67
4	0.167	6	−0.44	−0.06	0.45

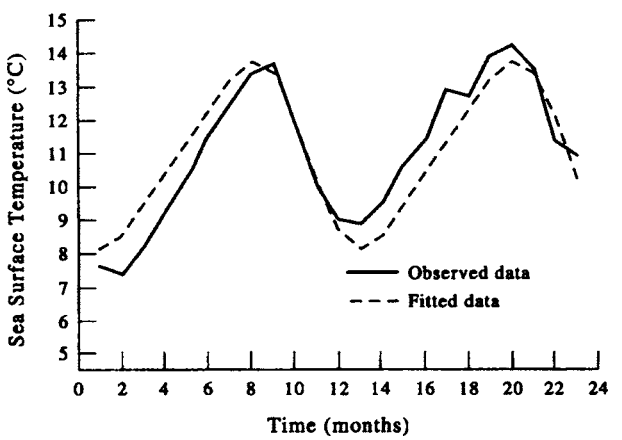


Figure 5.5.1. Monthly mean sea surface temperature (SST) record for Amphitrite Point on the west coast of Vancouver Island (see Table 5.2). The bold line is the original 24-month series. The dashed line is the SST time series obtained from a least-squares fit of the annual (12 month) and semi-annual (six month) cycles to the mean-removed data (see Table 5.3).

while that for the fitted record is 1.91°C. For this short segment of the data record, the two constituents account for 91.7% of the total variance.

5.5.3 Harmonic analysis of tides

Harmonic analysis is most useful for the analysis and prediction of tide heights and tidal currents. The use of this technique for tides appears to have originated with Lord Kelvin (1824–1907) around 1867. Lord Kelvin (Sir William Thomson) is also credited with inventing the first tide-predicting machine, although the first practical use of such a device was not made until several years later. A discussion of tidal harmonic analysis can be found in the *Admiralty Manual of Tides* (Doodson and Warburg, 1941) and Godin (1972). Definitive reports on the least-squares analysis of current and tide-height data were presented by Foreman (1977, 1978).

The least-squares harmonic analysis method has a variety of attractive features. It permits resolution of several hundred tidal constituents of which 45 are typically astronomical in origin and identified with a specific frequency in the tidal potential. The remaining constituents include shallow water constituents associated with bottom frictional effects and nonlinear terms in the equations of motion as well as radiational constituents originating with atmospheric effects. Both scalar and vector time series can be analyzed, with processing of vector series such as current velocity considerably more complex than processing of scalar time series such as sea level and water temperature. If the record is not sufficiently long to permit the direct resolution of neighboring components in the diurnal and semidiurnal frequency bands, the analysis makes provision for the "inference" and subsequent inclusion of these components in the analysis. For example, in the case of the diurnal constituent, P_1, associated with the sun's declination, the phase and amplitude are obtained by lowering the resolution criterion (called the *Rayleigh criterion*) for the separation of frequencies until P_1 is just resolved. The amplitude ratio (amp P_1/amp K_1) and phase difference (phase P_1–phase K_1) relative to the readily resolved diurnal constituent K_1 can then be calculated and used to calculate the P_1 constituent for the original record. Equally importantly, the method allows for gaps in the time series by ignoring those times for which there are no data. Major features of the least-squares optimization procedure for tidal analysis are outlined below.

The aim of least-squares analysis is to estimate the tidal harmonic constituent amplitudes and phases which can then be used for long-term tidal predictions. The commonly used sampling interval for tidal analysis is 1 h, so that even data collected at shorter time intervals are usually averaged to 1 h intervals for standard analysis packages. Records must have a minimum length of 13 h in order that they incorporate at least one cycle of the M_2 tidal frequency (period, 12.42 h). The mean component Z_0 is also included. As the length of the record is increased, additional constituents can be added to the analysis. (As noted in Chapter 1, our ability to resolve adjacent frequencies improves with the length of the time series. Aside from the degree of noise in the data, the main factor limiting the number of derived tidal constituents is the length of the record.) For example, the K_1 constituent (period, 23.93 h) can be adequately determined once the record length exceeds 24 h, although less reliable estimates can be made for shorter record lengths. The criteria for deciding which constituents can be included is discussed in the next section. In essence, inclusion requires that the difference in frequency, Δf, between a given constituent and its so-

called *Rayleigh reference constituent* be greater than the fundamental frequency for the record; i.e. $\Delta f \geq f_1 = 1/T$(see following discussion).

5.5.4 Choice of constituents

The least-squares method can be applied to any combination of tidal frequencies. However, the rational approach is to pick the allowable frequencies on the basis of two factors: (1) their relative contribution to the tide-generating potential; and (2) their resolvability in relation to a neighboring principal tidal constituent. In other words, the constituent should be one that makes a significant contribution to the tide-generating force and the record should be of sufficient duration to permit accurate separation of neighboring frequencies. Consideration should also be given to the required computational time, which increases roughly as the square of the number of constituents used in the analysis. Due to noise limitations, the amplitudes of many constituents are too small to be adequately resolved by most oceanic data sets.

To determine whether a specific constituent should be included in the tidal analysis, the frequency f_m of the constituent is compared to the frequency of the neighboring Rayleigh comparison constituent, f_R. The constituent can be included provided

$$|f_m - f_R|T = |\Delta f|T > R \tag{5.5.16}$$

where T is the record length and R is typically equal to unity (depending on background noise). In effect, equation (5.5.16) states that f_m should be included if f_R is an included frequency *and* the ratio of the frequency difference Δf to the fundamental frequency $f_1 = 1/T$ is greater than unity. This implies that the fundamental frequency, which corresponds to the best resolution (separation) achievable on the frequency axis, is less than the frequency separation between constituents. Values of $R < 1$ are permitted in the least-squares program to allow for approximate estimates of neighboring tidal frequencies for record lengths T shorter than $1/\Delta f$. Obviously, the longer the record, the more constituents are permitted.

The choice of f_R is determined by the hierarchy of constituents within the tidal band of interest and level of noise in the observations. The hierarchy is in turn based on the contribution a particular constituent makes to the equilibrium tide, with the largest contribution usually coming from the M_2 tidal constituent (Cartwright and Edden, 1973). For the major contributors to the equilibrium tide, the magnitude ratios relative to M_2 in descending order are: $K_1/M_2 = 0.584$, $S_2/M_2 = 0.465$, and $O_1/M_2 = 0.415$. Depending on the level of noise in the observations, the principal semidiurnal constituent M_2 (0.0805 cph) and the record mean Z_0 can be determined for records longer than about 13 h duration, while the principal diurnal component K_1 (0.0418 cph) can be determined for records longer than about 24 h. As a rough guide, separation of the next most significant semidiurnal constituent S_2 (0.0833 cph) from the principal component M_2 requires a record length $T > 1/|f(M_2) - f(S_2)| =$ 355 h (14.7 days). Similarly, separation of the next most significant diurnal constituent, O_1 (0.0387 cph), from the principal component, K_1, requires an approximate record length $T > 1/|f(K_1) - f(O_1)| = 328$ h (13.7 days). The frequencies $f(K_1)$ and $f(O_1)$ then become the Rayleigh comparison frequencies for other neighboring tidal constituents in the diurnal band while the frequencies $f(M_2)$ and $f(S_2)$ become the comparison frequencies for neighboring frequencies in the semidiurnal band. Extension of this procedure to longer and longer records eventually

encompasses all the significant tidal constituents within the diurnal and semidiurnal bands. The first long-term constituent to be included in the analysis is the lunar–solar fortnightly cycle M_{sf} (0.00282 cph), requiring an approximate record duration $T >$ 14.8 days, followed by the lunar monthly constituent M_m (0.00151 cph), duration $T >$ 31.8 days, and the lunar fortnightly cycle M_f (0.00305 cph), $T >$ 182.6 days. These record length requirements are based on stochastic processes; shorter records can be used for deterministic processes such as tides provided that noise levels are low. Thus, in all cases, shorter record lengths can be used if the data are highly noise free. By the same token, longer records are often needed to resolve the longer period tides because of contamination from atmospheric effects.

A summary of the required record lengths for inclusion of the more important constituents is provided in Tables 5.5–5.7 together with a comparison of the constituents tidal potential magnitude relative to that of the principal component in the frequency band. Where possible, a candidate constituent is compared to the particular neighboring constituent which has already been selected and is nearest in frequency.

5.5.5 A computational example for tides

As a simple example of the least-squares method of harmonic tidal analysis, consider the 32-hourly sea-level heights measured at Tofino, British Columbia during 10–11 September 1986 (Table 5.8). As indicated by Tables 5.5 and 5.6, we can at most resolve the K_1 and M_2 constituents. This problem is similar to that considered in Section 5.5.2 where we used the least-squares technique to fit the annual and semi-annual components to a 24-month record of sea surface temperature. Following the analysis in that section, the various matrices are written in terms of a mean component plus the contributions from the K_1 and M_2 frequencies, $f(K_1)=$ 0.0418 cph and $f(M_2)=$ 0.0805 cph, respectively. From (5.5.8) and (5.5.9), we find

$$\mathbf{D} = \begin{pmatrix} N & c_1 & c_2 & s & s_2 \\ c_1 & cc_{11} & cc_{12} & cs_{11} & cs_{12} \\ c_2 & cc_{21} & cc_{22} & cs_{21} & cs_{22} \\ s_1 & sc_{11} & sc_{12} & ss_{11} & ss_{12} \\ s_2 & sc_{21} & sc_{22} & ss_{21} & ss_{22} \end{pmatrix} \tag{5.5.17}$$

$$= \begin{pmatrix} 32 & 2.476 & -1.836 & 6.183 & 3.420 \\ 2.476 & 14.809 & 1.450 & 1.136 & 2.117 \\ -1.836 & 1.450 & 16.263 & -2.197 & 0.397 \\ 6.183 & 1.136 & -2.1 & 17.191 & 2.163 \\ 3.420 & 2.117 & 0.397 & 2.163 & 15.737 \end{pmatrix} \tag{5.5.18}$$

and from (5.5.9) and (5.5.11)

$$\mathbf{y} = \begin{pmatrix} yc_0 \\ yc_1 \\ yc_2 \\ ys_1 \\ ys_2 \end{pmatrix} = \begin{pmatrix} 57.640 \\ 6.514 \\ 6.138 \\ -0.199 \\ -3.335 \end{pmatrix} \tag{5.5.19}$$

Table 5.5. Record lengths to resolve main tidal constituents in the semidiurnal tidal band assuming a Rayleigh coefficient R = 1. Also listed are the comparison constituents and ratios of tidal potential to that of the principal semidiurnal constituent M_2

Tidal constituent	Frequency (cph)	Comparison constituent	Magnitude ratio	Record length (h)
M_2 (principal lunar)	0.0805	–	1	13
S_2 (principal solar)	0.0833	M_2	0.465	355
N_2 (larger lunar elliptic)	0.0790	M_2	0.192	662
K_2 (luni-solar)	0.0836	S_2	0.029	4383

Table 5.6. Record lengths to resolve main tidal constituents in the diurnal tidal band assuming a Rayleigh coefficient R = 1. Also listed are the comparison constituents and ratios of tidal potential to that of the principal semidiurnal constituent, M_2

Tidal constituent	Frequency (cph)	Comparison constituent	Magnitude ratio	Record length (h)
K_1 (luni-solar)	0.0418	–	0.584	24
O_1 (principal lunar)	0.0387	K_1	0.415	328
P_1 (principal solar)	0.0416	K_1	0.193	4383
Q_1	0.0372	O_1	0.079	662

Table 5.7. Record lengths to resolve main tidal constituents in the long-period tidal band assuming a Rayleigh coefficient R = 1. Also listed are the comparison constituents and ratios of tidal potential to that of the principal semidiurnal constituent, M_2

Tidal constituent	Frequency (cph)	Comparison constituent	Magnitude ratio	Record length (h)
M_{sf} (mixed solar fortnightly)	0.002822	M_f	0.015	355
M_f (lunar fortnightly)	0.003050	–	0.172	4383
M_m (lunar monthly)	0.001512	M_{sm}	0.091	764
M_{sm} (solar monthly)	0.001310	–	0.017	4942
S_{sa} (solar semi-annual)	0.000228	S_a	0.080	4383
S_a (solar annual)	0.000114	–	0.013	8766

Table 5.8. Hourly values of sea-level height (SLH) measured at Tofino, British Columbia (49°09.0′ N, 125°54.0′ W) on the west coast of Canada starting 10 September 1986. Heights are in meters above the local datum

n	1	2	3	4	2	6	7	8	9	10	11
SLH	1.97	1.46	0.98	0.73	0.67	0.82	1.15	1.58	2.00	2.33	2.48
n	12	13	14	15	16	17	18	19	20	21	22
SLH	2.43	2.25	2.02	1.82	1.72	1.75	1.91	2.22	2.54	2.87	3.10
n	23	24	25	26	27	28	29	30	31	32	
SLH	3.15	2.94	2.57	2.06	1.56	1.13	0.84	0.73	0.79	1.07	

where the elements of $\mathbf{D}$ and $\mathbf{y}$ have units of meters. The solution $\mathbf{z} = \mathbf{D}^{-1}\mathbf{y}$ is the vector

$$\mathbf{z} = \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} 1.992\,m \\ 0.186\,m \\ 0.523\,m \\ -0.574\,m \\ -0.604\,m \end{pmatrix} \tag{5.5.20}$$

The results are summarized in Table 5.9. A plot of the original sea-level data and the fitted sea-level curve are presented in Figure 5.5.2. The standard deviation for the original record is 0.741 m while that for the fitted record is 0.736 m. For this short segment of the data record, the sum of the two tidal constituents accounts for over 99% of the total variance in the record. As a comparison, we have used the full analysis package without inference to analyze 29 days of the Tofino sea-level record beginning at 2000 on 10 September 1986. The program finds a total of 30 constituents, including the mean, Z_0, with the sum of the tidal constituents accounting for 98% of the original variance in the signal. The record mean for the month is 2.05 m, and the K_1 and M_2 constituents have amplitudes of 0.286 and 0.986 m, respectively. As expected, these are quite different to the values derived on only 32 h of data (Table 5.9). Phases for the two constituents for the 29-day records are 122.0° and 12.5° compared with 107.9° and 130.9° for the same two constituents based on the 32-h records (angles in both cases are measured counterclockwise from the positive x axis).

Table 5.9. Least-squares estimates of the amplitude and phase of the K_1 and M_2 tidal constituents for the 32-h Tofino sea level starting at 2000, 10 September 1986. The mean is $\frac{1}{2}A_0$. The last column, C'_q, gives the constituent amplitudes for a more extensive analysis that used a 29-day (685 h) data segment that had the same start time as the 32-h segment used to derive C_q

q	Frequency (cph)	Period (h)	A_q (m)	B_q (m)	C_q (m)	C'_q (m)
0	–	–	3.984	0	3.984	4.100
1	0.042	24	0.186	−0.574	0.365	0.286
2	0.081	12	0.523	−0.604	0.638	0.986

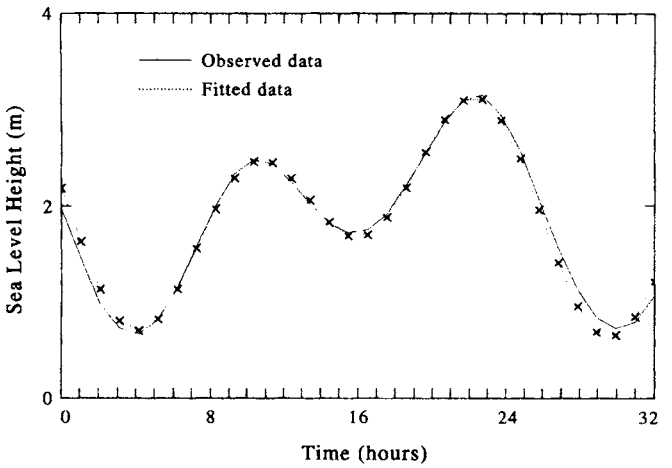


Figure 5.5.2. Hourly sea-level height (SLH) recorded at Tofino on the west coast of Vancouver Island (see Table 5.7). The solid line is the original 32-h series; the dotted line is the SLH series obtained from a least-squares fit of the main diurnal (K_1, 0.042 cph) and main semidiurnal (M_2, 0.081 cph) tidal frequencies to the mean-removed data (see Table 5.8).

5.5.6 Complex demodulation

In many applications, we seek to determine how the signal characteristics at a specific frequency, ω, change throughout the duration of a time series. For example, we might ask how the amplitude, phase, and orientation of the semidiurnal tidal current ellipses at different depths at a mooring location change with time. Wave packets associated with passing internal tides would be revealed through rapid changes in ellipse parameters at the M_2 and/or S_2 frequencies. The method for determining the temporal change of a particular frequency component for a velocity or scalar time series is called *complex demodulation*.

A common technique for finding the demodulated signal is to fit the desired parameters to sequential segments of the data series using least-squares algorithms. The analysis requires that there be many more data points than frequency components and each segment must span at least one cycle of the frequency of interest. As with any least-squares analysis, the observations do not have to be at regular time intervals. Inputs to complex demodulation algorithms require specification of the start time of the first segment, the length of each segment, and the time between computation interval start times. Computation intervals may overlap, be end-to-end, or be interspersed with unused data. Following the least-squares analysis described under the section on harmonic analysis, the time increment between each estimate can be as short as one time step, Δt, thereby providing the maximum number of estimates for a given segment length, or as long as the entire record, thereby yielding a single estimate of the signal parameters.

For each segment of current velocity data, the fluctuating component of velocity at frequency ω can be expressed as

$$\begin{aligned}\mathbf{u}(t) - \overline{\mathbf{u}(t)} &= \left[u(t) - \overline{u(t)}\right] + \mathrm{i}\left[v(t) - \overline{v(t)}\right] \\ &= A^+ \exp\left[\mathrm{i}(\omega t + \varepsilon^+)\right] + A^- \exp\left[-\mathrm{i}(\omega t + \varepsilon^-)\right]\end{aligned} \tag{5.5.21}$$

where $\overline{u(t)}$, $\overline{v(t)}$ are mean components of the velocity, and (A^+, A^-) are the amplitudes and $(\varepsilon^+, \varepsilon^-)$ the phases of the counterclockwise (+) and clockwise (−) rotating components. Data are at times t_k $(k = 1, \ldots, N)$ and solutions are found from the matrix equation

$$\mathbf{z} = \mathbf{D}^{-1}\mathbf{y} \tag{5.5.22}$$

where

$$\mathbf{y} = \begin{pmatrix} u(t_1) \\ u(t_2) \\ \cdots \\ u(t_n) \\ v(t_1) \\ \cdots \\ v(t_n) \end{pmatrix}; \quad \mathbf{z} = \begin{pmatrix} A^+ \cos(\varepsilon^+) \\ A^+ \sin(\varepsilon^+) \\ A^- \cos(\varepsilon^-) \\ A^- \sin(\varepsilon^-) \end{pmatrix} \equiv \begin{pmatrix} ACP \\ ASP \\ ACM \\ ASM \end{pmatrix} \tag{5.5.22a}$$

and

$$\mathbf{D} = \begin{pmatrix} \cos(\omega t_1) & -\sin(\omega t_1) & \cos(\omega t_1) & \sin(\omega t_1) \\ \cos(\omega t_2) & -\sin(\omega t_2) & \cos(\omega t_2) & \sin(\omega t_2) \\ \cdots & \cdots & \cdots & \cdots \\ \cos(\omega t_n) & -\sin(\omega t_n) & \cos(\omega t_n) & \sin(\omega t_n) \\ \sin(\omega t_1) & \cos(\omega t_1) & -\sin(\omega t_1) & \cos(\omega t_1) \\ \cdots & \cdots & \cdots & \cdots \\ \sin(\omega t_n) & \cos(\omega t_n) & -\sin(\omega t_n) & \cos(\omega t_n) \end{pmatrix} \tag{5.5.22b}$$

Once the elements of $\mathbf{z}$ are found from the least-squares solution to the matrix equation (for example, using IMSL routine LLSQAR), we can find the various ellipse parameters from

$$A^+ = (ASP^2 + ACP^2)^{1/2}; \quad A^- = (ASM^2 + ACM^2)^{1/2} \tag{5.5.23a}$$

$$\tan(\varepsilon^+) = \frac{ASP}{ACP}; \quad \tan(\varepsilon^-) = \frac{ASM}{ACM} \tag{5.5.23b}$$

For example, we could obtain the demodulated current amplitude and phase for near-inertial motions observed at a mid-latitude mooring by setting $\omega = 2\Omega \sin\theta$ and obtaining least-squares solutions for a series of adjoining 24-h segments with no overlap (here, Ω is the angular earth rotation rate and θ is latitude). For the least-squares technique to be applicable, data would need to sampled at roughly hourly intervals so that there were more data points per segment than parameters being estimated. Equatorward of $\theta = \pm 30°$ the period of inertial motions exceeds 24 h and the lengths of individual segments must be increased accordingly. Complex demodulation also can be used to examine inertial motions in Lagrangian-type data. In Figure 5.5.3(a), we have plotted the original and demodulated positions of a satellite-tracked drifter launched in the Canadian Arctic in the fall of 1988. The time series covers 60 days and was analyzed using overlapping 24-h subsections with the assumption that displacements occurred at the inertial period of 12.73 h for 70°N latitude. Figure 5.5.3(b) presents a detailed analysis of the trajectory record for the 20 days ending 11 October when the buoy became trapped in growing sea ice. Note the intense inertial currents starting on 30 September, the prevalence of the clockwise

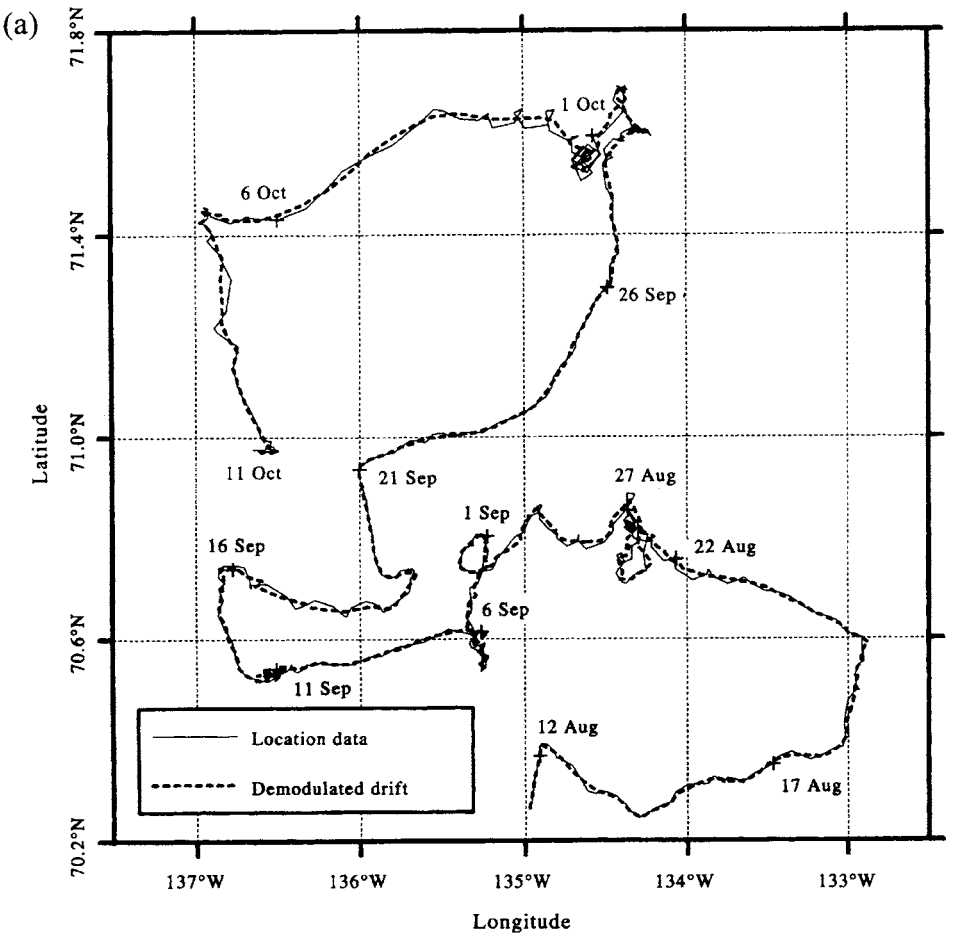


Figure 5.5.3. Complex demodulation at the inertial period of 12.73 h for the trajectory of a satellite-tracked drifter deployed in the Beaufort Sea in August 1988. (a) Original (solid line) and demodulated version (dashed line) of the drifter track. (Courtesy of Humfrey Melling.)

component of rotation, and the roughly −6.4° per day drift in phase of the clockwise component of the current due to the changing latitude of the drifter relative to the reference latitude of 70°N.

5.6 SPECTRAL ANALYSIS

Spectral analysis is used to partition the variance of a time series as a function of frequency. For stochastic time series such as wind waves, contributions from the different frequency components are measured in terms of the *power spectral density* (PSD). For deterministic waveforms such as surface tides, either the PSD or the *energy spectral density* (ESD) can be used. Here, power is defined as energy per unit time. The need for two different spectral definitions lies in the boundedness of the integral of signal variance for increasing record length. In practice, the term *spectrum* is applied to all spectral functions including commonly used terms such as autospectrum and

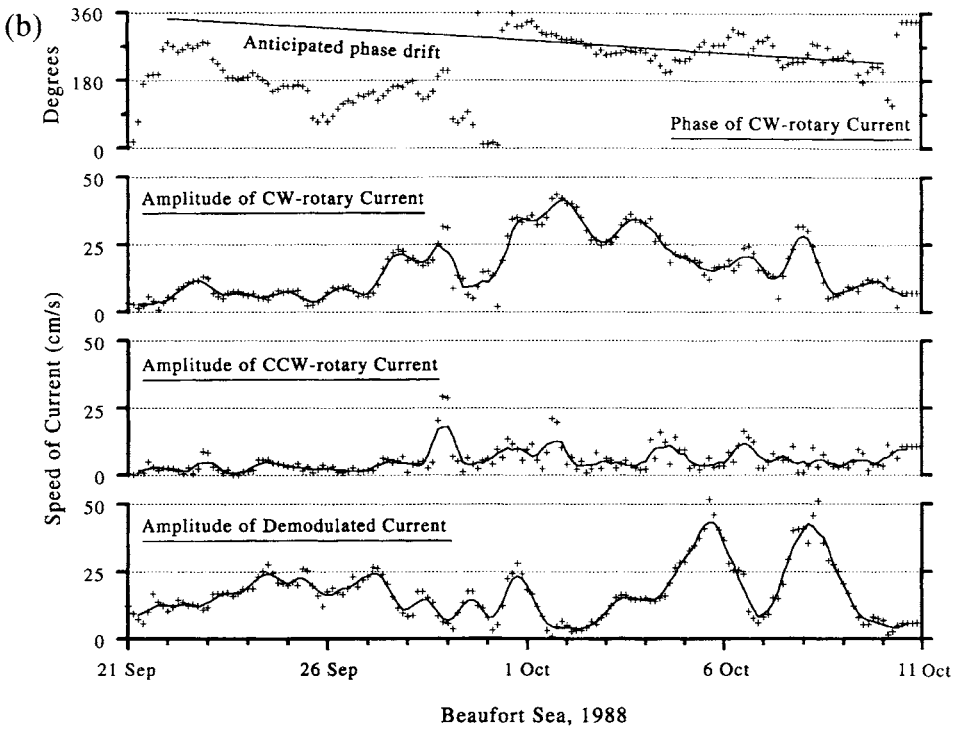


Figure 5.5.3. Complex demodulation at the inertial period of 12.73 h for the trajectory of a satellite-tracked drifter deployed in the Beaufort Sea in August 1988. (b) Parameters of the demodulation over a 20-day period of strong inertial motions. Top panel: phase of the clockwise (CW) rotary component (degrees). Remaining panels: amplitudes of the CW rotary, CCW rotary, and speed of the demodulated current. (Courtesy of Humfrey Melling.)

power spectrum. The term *cross-spectrum* is reserved for the "shared" power between two coincident time series. We also distinguish between *nonparametric* and *parametric* spectral methods. Nonparametric methods, which are based on conventional Fourier transforms, are not data-specific while parametric techniques are data-specific and assign a predetermined model to the time series. In general, we use parametric methods for short time series (few cycles of the oscillations of interest) and non-parametric methods for long time series (many cycles of oscillations of interest).

The word spectrum is a carry-over from optics. The colors red, white, and blue of the electromagnetic spectrum are often used to describe the frequency distribution of oceanographic spectra. A spectrum whose spectral density decreases with increasing frequency is called a "red" spectrum, by analogy to visible light where red corresponds to longer wavelengths (lower frequencies). Similarly, a spectrum whose magnitude increases with frequency is called a "blue" spectrum. A "white" spectrum is one in which the spectral constituents have near-equal amplitude throughout the frequency range. In the ocean, long-period variability (periods greater than several days) tend to have red spectra while instrument noise tends to have white spectra. Blue spectra are confined to certain frequency bands such as the low-frequency portion of wind–wave spectra and within the weather band ($2 <$ period < 10 days) for deep wind-generated currents.

In the days before modern computers it was customary to compute the spectrum of discrete oceanic data from the Fourier transform of the autocorrelation function using

a small number of lag intervals, or "lags". First formalized by Blackman and Tukey (1958), the autocorrelation method lacks the wide range of optional improvements to the computations and generalized "tinkering" permitted by more modern techniques. From a historical perspective, the autocorrelation approach has importance for the direct mathematical link it provides through the Wiener–Khinchin relations that link variance functions in the time domain to those in the frequency domain. Today, it is the spectral *periodogram* generated using the fast Fourier transform (FFT) or the Singleton Fourier transform that is most commonly used to estimate oceanic spectra.

Other methods have been developed over the years as a result of fundamental performance limitations with the periodogram method. These limitations are: (1) restricted frequency resolution when distinguishing between two or more signals, with frequency resolution dictated by the available record length independent of the characteristics of the data or its signal-to-noise ratio (SNR); (2) energy "leakage" between the main lobe of a spectral estimate and adjacent side-lobes, with a resulting distortion and smearing of the spectral estimates, suppression of weak signals, and the need to use smoothing windows; (3) an inability to adequately determine the spectral content of short time series; and (4) an inability to adjust to rapid changes in signal amplitude or phase. Other techniques, such as the maximum entropy method (best suited to short time series) and the wavelet transform (best suited to event-like signals), are addressed in this chapter.

Fundamental concepts: Several basic concepts are woven into the fabric of this chapter. First of all, the sample data we collect are subsets of either stochastic or deterministic processes. Deterministic processes are predictable, stochastic ones are not. Secondly, the very act of sampling to generate a time series of finite duration is analogous to viewing an infinitely long time series through a narrow "window" in the shape of a rectangular box-car function (Figure 5.6.1a). The characteristics of this window in the frequency domain can severely distort the frequency content of the original data series from which the sample has been drawn. As illustrated by Figure 5.6.1(b), the sampling process results in spectral energy being "rippled" away from one frequency (the central lobe of the response function) to a wide number range of adjacent frequencies. The large side-lobes of the rectangular window are responsible for the leakage of spectral energy from the central frequency to nearby frequencies.

A third point is that the spectra of random processes are themselves random processes. Therefore, if we are to determine the frequency content of a data series with some degree of statistical reliability (i.e. to be able to put confidence intervals on spectral peaks), we need to precondition the time series and average the raw periodogram estimates. Averaging can be done in the time domain by using specially designed windows or in the frequency domain by averaging together adjacent spectral estimates. Windows (which are discussed in detail in Section 5.6.6) suppress Gibbs' phenomenon associated with finite length data series and enable us to increase the number of *degrees of freedom* used in each spectral estimate. (Here, the term "degrees of freedom" refers to the number of statistically independent variables or values used in a particular estimate.) We can also improve spectral estimates by partitioning a time series into a series of segments and then conducting spectral analysis on the separate pieces. Spectral values in each frequency band for each piece are then averaged as a block to improve statistical reliability. The penalty for doing this is a loss in frequency resolution. The alternative—calculating a single periodogram and then smoothing in the frequency domain—suffers the same loss of frequency resolution for a smoothing that gives the same degrees of freedom.

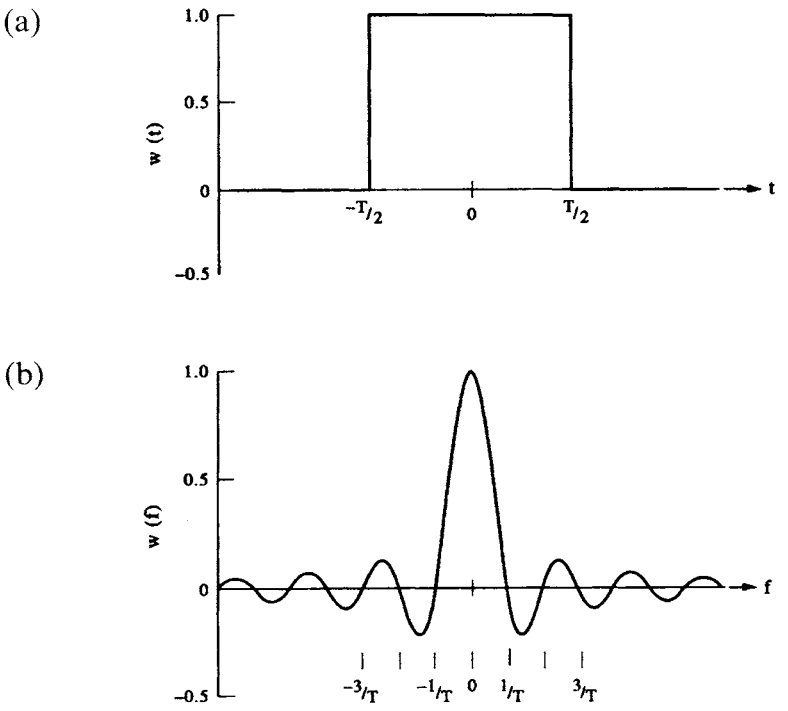


Figure 5.6.1. The box-car (rectangular) window which creates a sample time series from a "long" time series. (a) The box-car window in the time (t) domain. Here, $w(t) = 1, -T/2 \leq t \leq T/2$, *and* $w = 0$ *otherwise. (b) Frequency (f) response of the box-car window in (a). The central lobe straddles each spectral (frequency) component within the time series and has a width* $\Delta f = 2/T$. *Zeros occur at* $f = \pm m/T$, *where* $m = 1, 2, \ldots$.

Regardless of which averaging approach we choose, the results will be tantamount to viewing the data through another window in the frequency domain. Any smoothing window used to improve the reliability of the spectral estimates will again distort the results and impose structure on the data, such as periodic behavior, when no such structure may exist in the original time series. In addition, conventional methods make the implicit assumption that the unobserved data or correlation lag-values situated outside the measurement interval are zero, which is generally not the case. The smoothing window results in smeared spectral estimates. The more modern parametric methods allow us to make more realistic assumptions about the nature of the process outside the measurement interval, other than to assume it is zero or cyclic. This eliminates the need for window functions. The improvement over conventional FFT spectral estimates can be quite dramatic, especially for short records. However, even then, there remain pitfalls which have tended to detract from the usefulness of these methods to oceanography. Each new method has its own advantages and disadvantages that must be weighed in context of the particular data set and the way it has been collected. For time series with low signal-to-noise ratio (SNR), most of the modern methods are no better than the conventional FFT approach.

Means and trends: Prior to spectral analysis, the record mean and trend are generally removed from any time series (Figure 5.6.2). Unless stated otherwise, we will assume that the time series $y(t)$ we wish to process has the form $y'(t) = y(t) - \overline{y(t)}$ where $\overline{y(t)} = y_o + \alpha t$ is the mean value and αt is the linear trend (y_o and α are constants). If

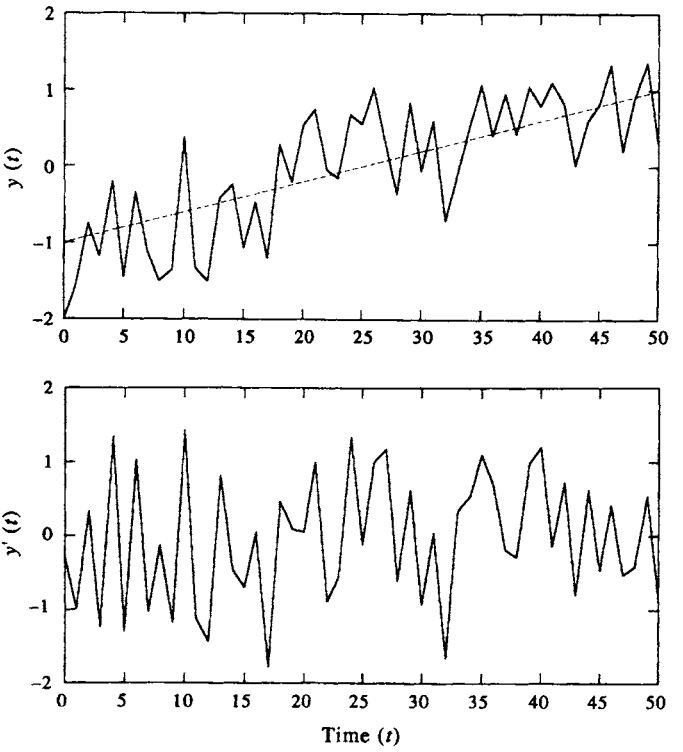


Figure 5.6.2. Mean and trend removal for an artificial time series y(t). Here, $y_o = -1.0$, trend $\alpha = 0.025$ and the fluctuating component, y', was obtained using a uniformly distributed random number generator. (a) Original time series, showing the linear trend; (b) Time series with the mean and linear trend removed.

the mean and trend are not removed prior to spectral analysis, they can distort the low-frequency components of the spectrum. Packaged spectral programs often include record mean and linear trend removal as part of the data preconditioning. Nonlinear trends are more difficult to remove, especially since a single function may not be appropriate for the entire data domain. The latter may apply also to linear trends.

The mean value removed from a record is not always the average for the entire record. For example, to examine interannual variability in the monthly time series of sea-level height, $\eta(t_m)$, at Cristobal on the Caribbean end of the Panama Canal, Thomson *et al.* (1985) first calculated mean-monthly values $\overline{\eta(t_m)}_m$ for each month (e.g. the individual means for January, February, etc.). These mean monthly values, rather than the average value for the entire record, were then subtracted from the original data for the appropriate month to obtain monthly anomalies of sea level, $\eta'(t_m) = \eta(t_m) - \overline{\eta(t_m)}_m$. Trend removal was then applied to the monthly anomalies to obtain the final sea-level anomaly record. As a final comment, we note that certain records, such as those from moored near-surface transmissometers, will contain nonlinear trends that should be removed from the data record prior to spectral analysis. This must be done with care. Unless one has a justified physical model for a particular trend (including a linear trend), removal of the trend may itself add spurious frequency components to the de-trended signal.

5.6.1 Spectra of deterministic and stochastic processes

Time-series data can originate with deterministic or stochastic processes, or a mixture of the two. Turbulence arising from eddy-like motions generated by strong tidal currents in a narrow coastal channel provides an example of mixed deterministic and stochastic processes. To see the difference between the two types of processes in terms of conventional spectral estimation, consider the case of a continuous *deterministic* signal, $y(t)$. If the total signal energy, E, is finite

$$E = \int_{-\infty}^{\infty} |y(t)|^2 \, dt < \infty \tag{5.6.1}$$

then $y(t)$ is absolute-integrable over the entire domain and the Fourier transform $Y(f)$ of $y(t)$ exists. This leads to the standard transform pair

$$Y(f) = \int_{-\infty}^{\infty} y(t) e^{-i2\pi ft} \, dt \tag{5.6.2a}$$

$$y(t) = \int_{-\infty}^{\infty} Y(f) e^{i2\pi ft} \, df = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(f) e^{i\omega t} \, d\omega \tag{5.6.2b}$$

where $e^{\pm i2\pi ft} = \cos(2\pi ft) \pm i \sin(2\pi ft)$, f is the frequency in cycles per unit time, and $\omega = 2\pi f$ is the angular frequency in radians per unit time. The square of the modulus of the Fourier transform for all frequencies

$$S_E(f) = Y(f)Y^*(f) = |Y(f)|^2 \tag{5.6.3}$$

is then the energy spectral density (ESD), $S_E(f)$, of $y(t)$. (As usual, the asterisk denotes the complex conjugate.) To see equation (5.6.3) that is an energy density, we use Parseval's theorem

$$\int_{-\infty}^{\infty} |y(t)|^2 \, dt = \int_{-\infty}^{\infty} |Y(f)|^2 df \tag{5.6.4}$$

which states that the total energy, E, of the signal in the time domain is equal to the total energy, $\int S_E(f) \, df$, of the signal in the frequency domain. Thus, $S_E(f)$, is an energy density (energy per unit frequency) which, when multiplied by df, yields a measure of the total signal energy in the frequency band centered near frequency f. The "power" of a deterministic signal, E/T, is zero in the limit of very long time series ($T \to \infty$).

Now, suppose that $y(t)$ is a stationary *random* process rather than a deterministic waveform. Unlike the case for the finite energy deterministic signal, the total energy in the stochastic process is unbounded (the characteristics of the process remain unchanged over time) and functions of the form (5.6.2) do not exist. In other words, the Fourier transform method introduced earlier fails in the sense that the total energy, as defined by equation (5.6.1), does not decrease as the length of the time

series increases without bound. To get around this problem, we must deal with the frequency distribution of the signal *power* (the time average of energy or energy per unit time, E/T) which is a bounded function. The basis for spectral analysis of random processes is the autocorrelation function $R_{yy}(\tau) = E[y(t)y(t+\tau)]$. Using the Wiener–Khinchin relation, the power spectral density, $S(f)$, becomes

$$S(f) = \int_{-\infty}^{\infty} R_{yy}(\tau) e^{-i2\pi f\tau}\, d\tau \tag{5.6.5a}$$

For an ergodic random process, for which ensemble averages can be replaced by time averages, R_{yy} has the form

$$R_{yy}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} [y(t)y^*(t+\tau)]\, dt \tag{5.6.5b}$$

By definition, the energy and power spectral density functions quantify the signal variance per unit frequency. For example, in the case of a stationary random process, integration of $S(f)$ gives the relation

$$s^2 = \int_{f-\Delta f/2}^{f+\Delta f/2} S(f)\, df \tag{5.6.6}$$

where s^2 is the integrated signal variance in the narrow frequency range $\Delta f = [f - 1/2\Delta f, f + 1/2\Delta f]$. If we assume that the spectrum is nearly uniform over this frequency range, we find

$$S(f) \approx \frac{s^2}{\Delta f} \tag{5.6.7}$$

which defines the spectrum for a stochastic processes in terms of a power density, or variance per unit frequency. The product $S(f)\Delta f$ is the total signal variance within the frequency band Δf centered at frequency f.

At this point, there are several other basic concepts worth mentioning. First of all, a waveform whose autocorrelation function $R(\tau)$ attenuates slowly with time lag, τ, will have a narrow spectral distribution (Figure 5.6.3a) indicating that there are relatively few frequency components to destructively interfere with one another as τ increases from zero. In the limiting case of only one frequency component, f_a, we find $R(\tau) \approx \cos(2\pi f_a \Delta t)$ and Fourier *line spectra* appear at frequencies $\pm f_a$ (Figure 5.6.3b). Because they consist of near monotone signals, tidal motions are highly autocorrelated and produce sharp spectral lines. In contrast, a rapidly decaying autocorrelation function implies a broad spectral distribution (Figure 5.6.4a) and a large number of frequency components in the original waveform. In the limit $R(\tau) \to \delta(\tau)$ (Figure 5.6.4b), there is an infinite number of equal-amplitude frequency components in the waveform and the spectrum $S(f) \to$ constant (white spectrum).

Figure 5.6.5 provides an example of time-series data generated by the relation $y(k) = A \cdot \cos(2\pi nk/N) + \varepsilon(k)$, where $k = 0, \ldots, N$ is time in units of $\Delta t = 1$, $n/N\Delta t = 0.25$ is the frequency in units of Δt^{-1}, and $\varepsilon(k)$ is a random number between

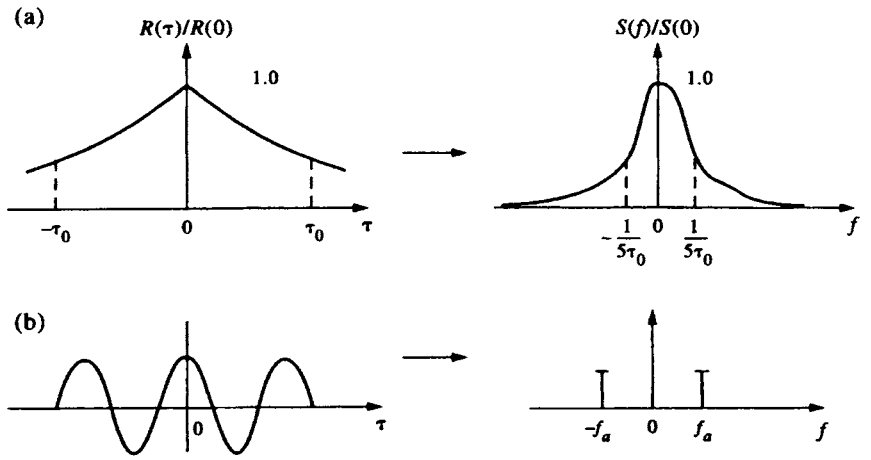


Figure 5.6.3. Examples of slowly decaying autocorrelation functions, $R(\tau)$, as a function of time lag, τ. Functions are normalized by their peak values. (a) The correlation function for a highly correlated signal leads to a relatively narrow power spectra density distribution, $S(f)$; (b) the case for autocorrelation $R(\tau) \approx \cos(2\pi f_a \Delta t)$ for a single frequency component, f_a, and corresponding line spectra at frequencies $\pm f_a$. (From Konyaev, 1990.)

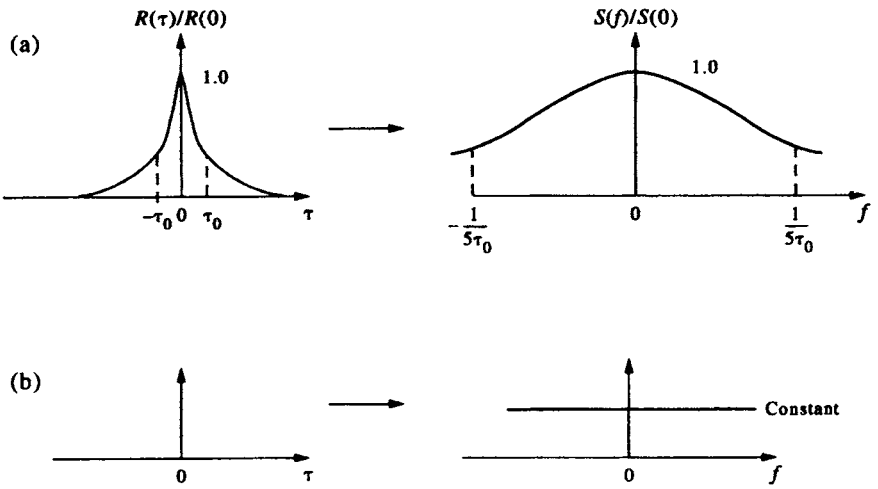


Figure 5.6.4. As for Figure 5.6.3 but for rapidly decaying autocorrelation functions, $R(\tau)$. (a) Correlation function for a weakly correlated signal leading to a broad power spectra density distribution. (b) The limiting case $R(\tau) \approx \delta(\tau)$ and the related spectrum $S(f)$ = constant (a white spectrum). (From Konyaev, 1990.)

−1 and +1. (We will often use this type of generic example rather than a specific example from the oceanographic literature. That way, readers can directly compare their computational results with ours. In the present case, if we set $\Delta t = 1$ day, then the time series $y(k)$ could represent east–west current velocity oscillations of a synoptic (three to 10-day) period associated with wind-forced motions (cf. Cannon and Thomson, 1996). Here, we set $A = 1$ and $\varepsilon(k) \neq 0$ for mostly deterministic data (Figure 5.6.5a) and $A = 0$ for random data (Figure 5.6.5b). In the analysis, the record has been padded with zeros up to time $k = 2N$. For the mostly deterministic case, the noise causes partial decorrelation of the signal with lag, but the spectral peak remains prominent.

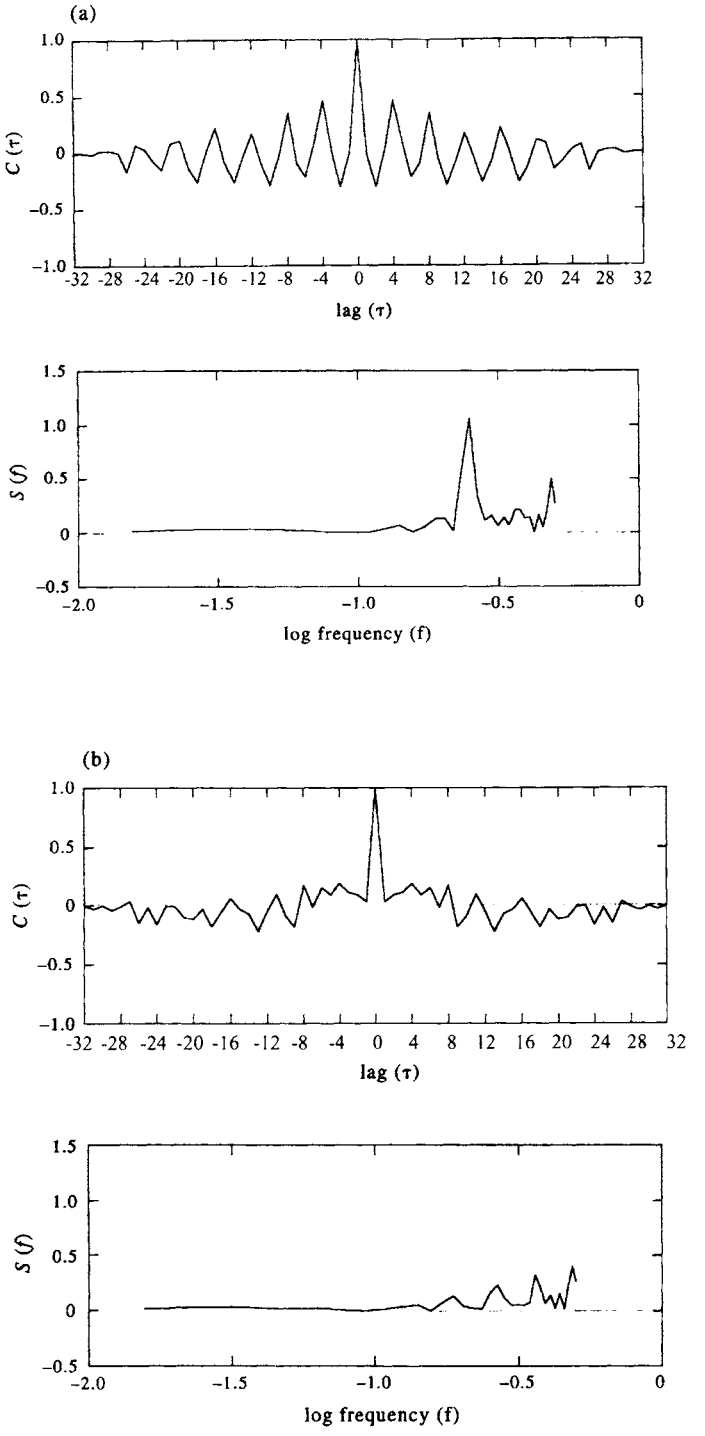


Figure 5.6.5. Autocovariance function $C(\tau)$ and corresponding spectrum $S(f)$ for the time series $y(k) = A\cos(2\pi nk/N) + \varepsilon(k)$; $k = 0, \ldots, N$, $\Delta t = 1$, $n/N = 0.25$ is the frequency, and $\varepsilon(k)$ is a random number between -1 and $+1$. (a) $C(\tau)$ and $S(f)$ for $A = 1$ and $\varepsilon \neq 0$ (mostly deterministic data); and (b) for $A = 0$ (purely random data). Records have been padded with zeros up to time $k = 2N = 32$.

For the purely random case, the spectrum resembles white noise but with isolated spectral peaks that one might mistake as originating with some physical process. The latter result is a good example of why we need to attach confidence limits to the peaks of spectral estimates (see Section 5.6.8).

5.6.2 Spectra of discrete series

Consider an infinitely long time series $y(t_n) = y_n$ sampled at equally spaced time increments $t_n = n\Delta t$, where Δt is the sampling interval and n is an integer, $-\infty < n < \infty$. From sampling theory, we know that a continuous representation of the discrete times series $y_s(t)$, can be represented as the product of the continuous time series $y(t)$ with an infinite set of delta functions, $\delta(t)$, such that

$$\begin{aligned} y_s(t) &= y(t) \sum_{n=-\infty}^{\infty} \delta(t - n\Delta t) \\ &= y(t) \frac{\Xi(t/\Delta t)}{\Delta t} \end{aligned} \tag{5.6.8a}$$

where Ξ is the "sampling function" and for which the Fourier transform is

$$\begin{aligned} Y(f) &= \int_{-\infty}^{\infty} \left[\sum_{n=-\infty}^{\infty} y(t)\delta(t - n\Delta t)\Delta t \right] e^{-i2\pi ft}\, dt \\ &= \Delta t \sum_{n=-\infty}^{\infty} y_n e^{-i2\pi ft} \end{aligned} \tag{5.6.8b}$$

In effect, the original time series is multiplied by a "picket fence" of delta functions $\Xi(t/\Delta t) \approx \sum_{n=-\infty}^{\infty} \delta(t - n\Delta t)$ which are zero everywhere except for the infinitesimal rectangular region occupied by each delta function (Figures 5.6.6a, b). Comparison of the above expression with equation (5.6.2) shows that retention of the time step Δt ensures conservation of the rectangular area in the two expressions as $\Delta t \to 0$. Provided that the time series $y(t)$ has a limited number of frequencies (i.e. is band-

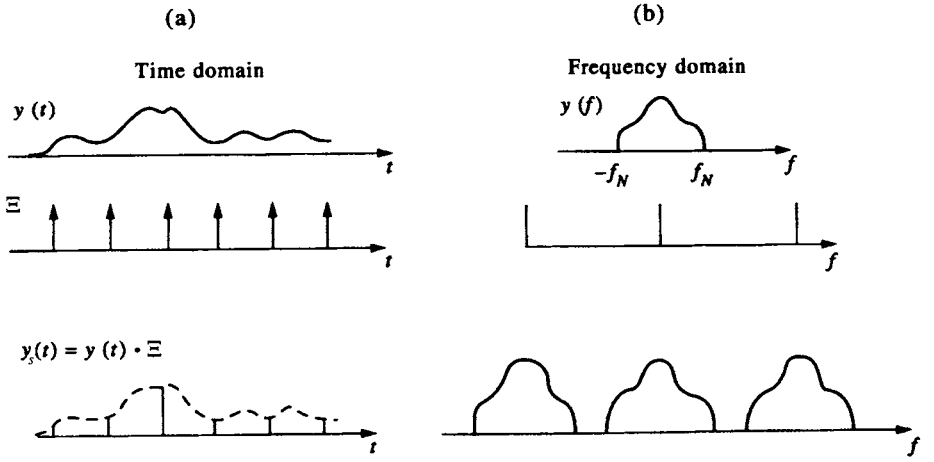


Figure 5.6.6. (a) A "picket fence" of delta functions $\delta(t - n\Delta t)$ used to generate a discrete data series from a continuous time series. (b) The Fourier transform (schematic only) of the different functions.

limited), whereby all frequencies are contained in the Nyquist interval

$$-f_N \leq f_k \leq f_N \tag{5.6.9}$$

in which $f_N = 1/(2\Delta t)$ is the Nyquist frequency, the energy spectral density

$$S_E(f) = |Y(f)|^2 \tag{5.6.10}$$

is identical to that for a continuous function. Conversely, if $Y(f) \neq 0$ for $|f| > f_N$ then the sampled and original times series do not have the same spectrum for $|f| < f_N$. The spectrum (5.6.10) obtained by Fourier analysis of discrete time series is called a *periodogram* spectral estimate, a term first coined by Schuster (1898) in a study of sunspot cycles.

Real oceanographic time-series data are discrete and have finite duration, $T = N\Delta t$. Returning to (5.6.8), this means that the summation is over a limited range $n = 1$ to N, and the spectral amplitude for the sample must be defined in terms of the discrete Fourier transform

$$\begin{aligned} Y_k &= \Delta t \sum_{n=1}^{N} y_n e^{-i2\pi f_k n \Delta t} \\ &= \Delta t \sum_{n=1}^{N} y_n e^{-i2\pi kn/N}; \quad f_k = k/N\Delta t, \; k = 0, \ldots, N \end{aligned} \tag{5.6.11}$$

The frequencies f_k are confined to the Nyquist interval, with positive frequencies, $0 \leq f_k \leq f_N$, corresponding to the range $k = 0, \ldots, N/2$ and negative frequencies, $-f_N \leq f_k \leq 0$, to the range $k = N/2, \ldots, N$. Since $f_{N-k} = f_k$, only the first $N/2$ Fourier transform values are unique. Specifically, $Y_k = Y_{N-k}$ so that we will generally confine our attention to the positive interval only.

The inverse Fourier transform is defined as

$$y_n = \frac{1}{N\Delta t} \sum_{k=0}^{N-1} Y_k e^{i2\pi kn/N}, \quad n = 1, \ldots, N \tag{5.6.12}$$

As indicated by equation (5.6.11), the Fourier transforms, Y_k, are specified for the discretized frequencies f_k, where $f_k = kf_1$ and $f_1 = 1/N\Delta t = 1/T$ characterizes both the fundamental frequency and the bandwidth, Δf, for the time series. The energy spectral density for a discrete, finite-duration time series is then

$$S_E(f_k) = |Y_k|^2, \quad k = 0, \ldots, N-1 \tag{5.6.13}$$

and Parseval's energy conservation theorem (5.6.4) becomes

$$\Delta t \sum_{n=1}^{N} |y_n|^2 = \Delta f \sum_{k=0}^{N-1} |Y_k|^2$$

where we have used $\Delta f = 1/(N\Delta t)$. A plot of $|Y_k|^2$ versus frequency, f_k, gives the discrete form of the periodogram spectral estimate.

Any geophysical data set we collect is subject to discrete sampling and windowing. As noted earlier, a time series of geophysical data, $y(t_n)$, sampled at time steps Δt can be considered the product of an infinitely long time series with a rectangular window which spans the duration ($T = N\Delta t$) of the measured data. The discrete spectrum $S(f_k)$ is the then the *convolution* of the true spectrum, $S(f)$, with the Fourier transform

of the rectangular window (Figure 5.6.1b). Since the window allows us to see only a segment of the infinite time series, the spectrum $S(f_k)$ provides a distorted picture of the actual underlying spectrum. This distortion, created during the Fourier transform of the rectangular window, consists of a broadening of the central lobe and leakage of power from the central lobe into the side-lobes. (The "ripples" on either side of the central lobe in Figure 5.6.1(b) are side lobes.) A further problem is that the function Y_k and its Fourier transform now become periodic with period N, although the original infinite time series $y(t)$, of which our sample data are a subset, may have been nonperiodic.

As noted in the previous section, the convergence of $|Y(f)|^2$ to $S(f)$ is smooth for deterministic functions in that the function $|Y'(f)|^2$, obtained by increasing the sample record length from T to T', would be a smoother version of $|Y(f)|^2$. For stochastic signals, the function $|Y'(f)|^2$ obtained from the longer time series (T') is just as erratic as the function for the shorter series. The sample spectra of a stochastic process do not converge in any statistical sense to a limiting value as T tends to infinity. Thus, the sample spectrum is not a consistent estimator in the sense that its PDF does not tend to cluster more closely about the true spectrum as the sample size increases. To show what we mean, consider the spectrum of a process consisting of $N = 400$ random, normally distributed deviates (Gaussian white noise) sampled at 1-s intervals. (True white noise is a mathematical construct and is as physically impossible as the spike of an impulse function.) The highest frequency we can hope to measure with these data is the Nyquist frequency, $f_N = 0.5$ cps. The spectra computed from 50 and then from 100 values of the fully white noise signal are presented in Figure 5.6.7(a). Also shown is the theoretical sample spectrum, corresponding to a uniform amplitude of 1.0. The shorter the sample used for the discrete spectral estimates, the greater the amplitude spikes in the power spectrum. This same tendency also is apparent in Table 5.6.1 which lists the means, variances, and mean square errors computed from various subsamples of the white noise signal. Here, mean square error (MSE) is defined as the variance plus bias of an estimator $\hat{y}(t)$ of the true signal $y(t)$; that is

$$\text{MSE} = E[(\hat{y} - y)^2] = V[\hat{y}] + B^2 \tag{5.6.14}$$

where $B = E[\hat{y}] - y$ is the bias of the estimator. The mean is lower in both the $N = 50$ and $N = 400$ cases while it is greater in the case where $N = 100$ and is exactly 1.0 for $N = 200$. The variance increases as N increases, as does the MSE. However, if this were a purely random discrete process (discrete white noise), the sample spectral estimator of the variance would be independent of the number of observations.

Now consider the spectrum of a second-order autoregressive process for a sample of $N = 400$ measured at 1-s increments (Figure 5.6.7b). (An autoregressive process of order p is one in which the present value of y depends on a linear combination of the

Table 5.6.1. Behavior of sample spectra of white noise as the record length is increased. (After Jenkins and Watts, 1968)

Record length (N)	50	100	200	400
Mean	0.85	1.07	1.00	0.95
Variance	0.630	0.777	0.886	0.826
Mean square error	0.652	0.782	0.886	0.828

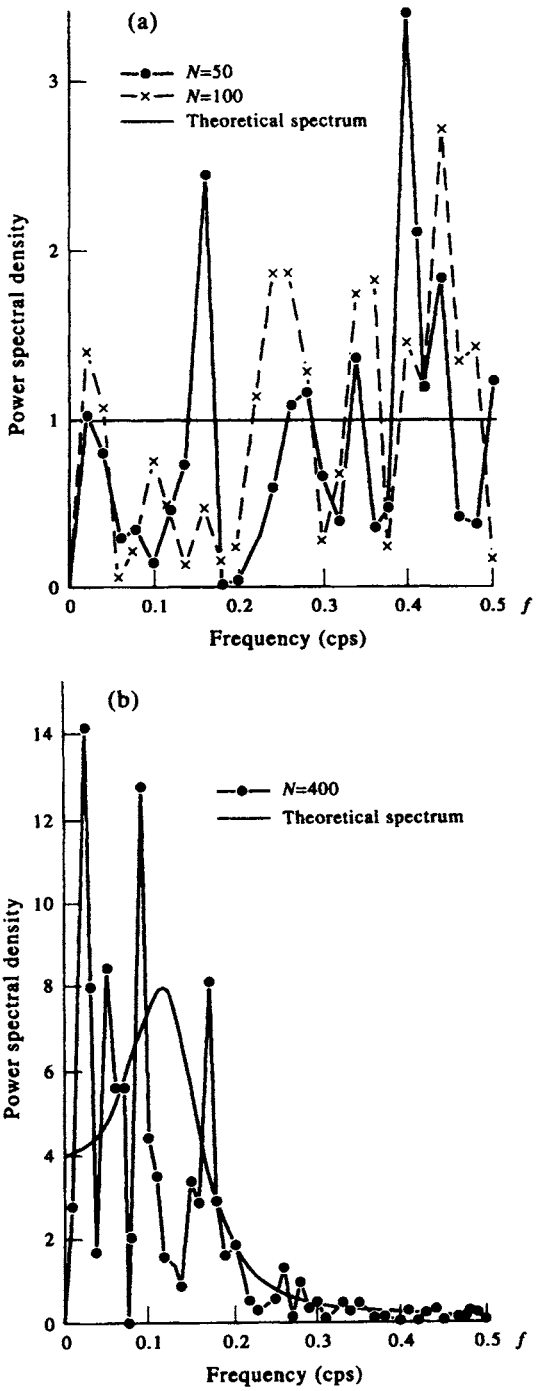


Figure 5.6.7. Power spectra of discrete signals and their theoretical values. Frequency in cycles per second (cps); spectra are in units of amplitude-squared/cps. (a) Power spectrum for the first half (N = 50) and full (N = 100) realization of a discrete normal white-noise process measured at 1-s intervals. (b) Power spectrum for one realization of a second-order autoregressive process of N = 400 values measured at 1-s increments. f_N = 0.5 cps is the Nyquist frequency and the maximum bandwidth of the spectral resolution $\Delta f = 1/N\Delta t = 0.0025$ /s. (From Jenkins and Watts, 1968.)

previous p values of y. See Section 5.7.2.) The Nyquist frequency is again 0.5 cps and the maximum bandwidth of the spectral resolution, $\Delta f = 1/N\Delta t$, is equal to 0.0025 cps. At the higher frequencies, the sample spectrum appears to be a good estimator of the theoretical spectrum (the smooth solid line), while for the lower frequencies there are large spikes in the sample spectrum which are not characteristic of the true spectrum. This misleading appearance is largely a consequence of the fact that the theoretical spectrum has most of its energy at the lower frequencies. In reality, the computed raw spectrum (i.e. with no smoothing) can fluctuate by 100% about the mean spectrum. The fluctuations are much smaller at higher frequencies simply because the actual spectral level is correspondingly smaller.

The basic reason why Fourier analysis breaks down when applied to real time series is that it is based on the assumption of fixed (stationary) amplitudes, frequencies, and phases. Stochastic series are instead characterized by random changes in frequency, amplitude, and phase. Thus, our treatment must be a statistical approach that makes it possible to accommodate these types of changes in our computation of the power spectrum.

5.6.3 Conventional spectral methods

The two spectral estimation techniques founded on Fourier transform operations are the indirect autocorrelation approach popularized by Blackman and Tukey in the 1950s and the direct periodogram approach presently favored by the oceanographic community. The fast Fourier transform (FFT) is the most common algorithm for determining the periodogram. The autocorrelation approach is mainly included for completeness. These methods fall into the category of nonparametric techniques which are defined independently of any specific time series. Parametric techniques, described later in this chapter, make assumptions about the variability of the time series and rely on the series for parameter determination.

The following sections first describe the two conventional spectral analysis methods without providing details on how to improve spectral estimates. We wish to first outline the procedures for calculating spectra before describing how to improve the statistical reliability of the spectral estimates. Once this is done, we give a thorough description of windowing, frequency-band averaging, and other spectral improvement techniques.

5.6.3.1 The autocorrelation method

In the Blackman–Tukey method, the autocovariance function, $C_{yy}(\tau)$ (which equals the autocorrelation function, $R_{yy}(\tau)$, if the record mean has been removed), is first computed as a function of lag, τ, and the Fourier transform of $C_{yy}(\tau)$ used to obtain the PSD as a function of frequency. An unbiased estimator for the autocovariance function for a data set consisting of N equally spaced values $\{y_1, y_2, \ldots, y_N\}$ is

$$C_{yy}(\tau_m; N-m) = \frac{1}{N-m}\sum_{n=1}^{N-m} y_n y_{n+m} \tag{5.6.15a}$$

where $m = 0, \ldots, M$ is the number of lags ($\tau_m = m\Delta t$) and $M < N$. In place of this estimator, some authors (cf. Kay and Marple, 1981) argue for the use of

$$C_{yy}(\tau_m; N) = \frac{1}{N}\sum_{n=1}^{N-m} y_n y_{n+m} \tag{5.6.15b}$$

which typically has a lower mean square error than $C_{yy}(\tau_m; N - m)$ for most finite data sets. Because $E[C_{yy}(\tau_m; N)] = [(N - m)/N]C_{yy}(\tau_m; N - m)$, the function $C_{yy}(\tau; N)$ is a biased estimator for the autocovariance function. Despite this, we will often use the relation (5.6.15b) for the autocovariance function since it yields a power spectral density (PSD) that is equivalent to the PSD obtained from the direct application of the FFT, as discussed in the next section. The weighting $(N - m)/N$ acts like a triangular (Bartlett) smoothing window to help reduce spectral leakage. We will use equation (5.6.15a) when we want a "stand-alone" unbiased estimator of the covariance function.

The one-sided power spectral density, G_k, for an autocovariance function with a total of M lags is found from the Fourier transform of the autocovariance function

$$G_k = 2\Delta t \sum_{m=0}^{M} C_{yy}(\tau_m) e^{-i2\pi km/M}, \quad k = 0, \ldots, (M/2) \tag{5.6.16a}$$

where $\tau_m = m\Delta t$ and $2\Delta t = 1/f_N$. Since $C_{yy}(\tau_m)$ is an even function, the spectrum of $\{y_n\}$ can be calculated from the cosine transform

$$G_k = 2\Delta t \left[C_{yy}(0) + 2\sum_{m=1}^{M} C_{yy}(\tau_m) \cos\frac{2\pi km}{N} \right], \quad k = 0, \ldots, (M/2) \tag{5.6.16b}$$

where $G_k = 2S_k$ is centered at positive frequencies $f_k = k/N\Delta t$ and the Nyquist interval $0 \le f_k \le f_N$ is divided into $N/2$ segments (N is even). For the two-sided spectrum, S_k, the first $(N/2) + 1$ frequencies are identical to those for the one-sided spectrum and correspond to positive frequencies in the range $0 \le f_k \le f_N$. The last $(N/2) - 1$ spectral values for the two-sided spectral density, defined for $k = (N/2) + 1, (N/2) + 2, \ldots, N - 1$, correspond to spectral density estimates for negative frequencies in the range $-f_N \le f_k \le 0$.

The solid line in Figure 5.6.8 shows spectra of monthly mean sea surface temperatures derived from the cosine transform using the Blackman–Tukey autocorrelation method for the version (5.6.15b) of the autocovariance function. The temperature data span the 36-month period from January 1982 to December 1984 for Amphitrite Point (Table 5.6.2). Since, in the next section, we wish to compare these spectra directly with those derived from the data series using a packaged FFT routine (the dashed curve in Figure 5.6.8), the lags were computed for the first 32 (2^5) points only, four fewer points than used in the Blackman–Tukey approach. In this case, extending the lag correlation beyond 10–20% of the data, as recommended earlier, is a necessity if we are to obtain reasonable estimates of the spectra. As expected, results reveal a strong spectral peak centered near, but not at, the annual frequency ($f = 1.0$ cycles per year = 0.083 cpmonth). There are too few data to enable us to accurately resolve the location of the frequency peak. In the present example, all spectral estimates are positive. However, the autocorrelation method can yield erroneous negative spectra for weak frequency components when there are gaps in the data record.

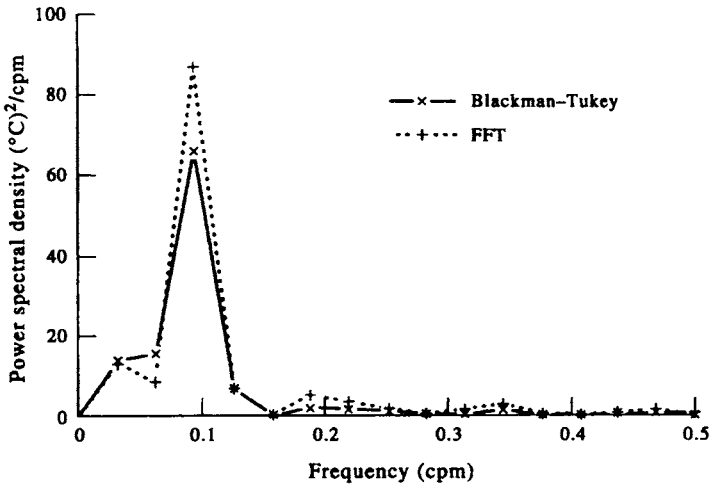


Figure 5.6.8. Spectra $(°C)^2$/cpm (cpm = cycles per month) versus frequency (per month) for monthly mean sea surface temperatures collected at a coastal station in the northeast Pacific for the period January 1982 to December 1984 (cf. Table 5.6.2). (a) The solid line is the unsmoothed spectrum from the Blackman–Tukey autocorrelation method (the cosine transform of the autocovariance function (5.6.15b)); dashed line is the unsmoothed spectrum from the FFT method based on the first 2^5 (= 32) data values. Spectral peaks span the annual period (f = 0.083 /month).

Table 5.6.2. Monthly mean sea surface temperatures SST (°C) at Amphitrite Point (48°55.16′ N, 125°32.17′ W) on the west coast of Canada for January 1982 through December 1984

Year 1982												
n	1	2	3	4	5	6	7	8	9	10	11	12
SST	7.6	7.4	8.2	9.2	10.2	11.5	12.4	13.4	13.7	11.8	10.1	9.0
Year 1983												
n	13	14	15	16	17	18	19	20	21	22	23	24
SST	8.9	9.5	10.6	11.4	12.9	12.7	13.9	14.2	13.5	11.4	10.9	8.1
Year 1984												
n	25	26	27	28	29	30	31	32	33	34	35	36
SST	7.9	8.4	9.3	9.9	11.0	11.1	12.6	14.0	13.0	11.7	9.8	8.0

We emphasize that the spectra in Figure 5.6.8 have been constructed without any averaging or windowing. This means that each spectral estimate has the minimum possible two degrees of freedom so that the error in each estimate is equal to the value of the estimate itself. Some form of averaging is needed if we are to place confidence limits on our spectra (see Sections 5.6.6 and 5.6.7). The two spectra are slightly different because the record used for the FFT method is shorter than that used for the autocovariance method.

5.6.3.2 The periodogram method

The preferred method for estimating the power spectral density of a discrete sample $\{y_1, y_2, \ldots, y_N\}$ is the direct or periodogram method. Instead of first calculating the autocorrelation function, the data are transformed directly to obtain the Fourier components $Y(f)$ using (5.6.11). To help avoid end effects (Gibbs' phenomenon) and

wrap-around problems, the original time series can be padded with $K \leq N$ zeros after the mean has been removed from the time series. The padding will also increase the frequency resolution of the periodogram (see Section 5.6.9). Although use of $K = N$ zeros is not recommended for computational reasons, it has one advantage: The N-lag covariance function obtained from the inverse Fourier transform of the $2N$-point power spectral density is identical to the N-lag covariance function (5.6.15b). As with the autocorrelation method, improvements in the statistical reliability of the spectral estimates would be provided by "windowing" the time series prior to spectral estimation or by averaging over the raw periodogram estimates over adjacent frequency bands (see Sections 5.6.6 and 5.6.7).

The two-sided power spectral (or autospectral) density for frequency f in the Nyquist interval $-1/(2\Delta t) \leq f \leq 1/(2\Delta t)$ and a padding of K zeros is

$$\begin{aligned} S_{yy}(f) &= \frac{1}{(N+K)\Delta t}\left|\Delta t \sum_{n=0}^{N+K-1} y_n e^{-i2\pi f n \Delta t}\right|^2 \\ &= \frac{1}{(N+K)\Delta t}|Y(f)|^2 \end{aligned} \tag{5.6.17a}$$

while the one-sided power spectral density for the positive frequency interval only, $0 \leq f \leq 1/(2\Delta t)$, is

$$G_{yy}(f) = 2S_{yy}(f) = \frac{2}{(N+K)\Delta t}|Y(f)|^2 \tag{5.6.17b}$$

Division by Δt transforms the energy spectral density of (5.6.13) into a power spectral density, $S_{yy}(f)$.

Evaluation of (5.6.17a) using the fast Fourier transform defines $Y(f)$ in terms of the discrete Fourier transform estimates, $Y(f_k) = Y_k$, where the f_k form a discrete set of $(N+K)/2$ equally spaced frequencies $f_k = \pm k/[(N+K)\Delta t]$, $k = 0, 1, \ldots, [(N+K)/2] - 1$ in the Nyquist interval, $-1/2\Delta t \leq f_k \leq 1/2\Delta t$. The case $k = 0$ represents the mean component. The two-sided PSD is then

$$S_{yy}(0) = \frac{1}{(N+K)\Delta t}|Y_0|^2, \quad k = 0$$

$$S_{yy}(f_k) = \frac{1}{(N+K)\Delta t}\left[|Y_k|^2 + |Y_{N+K-k}|^2\right], \quad k = 1, \ldots, \frac{(N+K)}{2} - 1 \tag{5.6.18a}$$

$$S_{yy}(f_N) = S_{yy}(f_{(N+K)/2-k}) = \frac{1}{(N+K)\Delta t}|Y_{(N+K)/2}|^2, \quad k = \frac{(N+K)}{2}$$

and the one-sided PSD is

$$G_{yy}(0) = \frac{1}{(N+K)\Delta t}|Y_0|^2, \quad k = 0$$

$$G_{yy}(f_k) = \frac{2}{(N+K)\Delta t}|Y_k|^2, \quad k = 1, \ldots, \frac{(N+K)}{2} - 1 \tag{5.6.18b}$$

$$G_{yy}(f_N) = G_{yy}(f_{(N+K)/2-k}) = \frac{1}{(N+K)\Delta t}|Y_{(N+K)/2}|^2, \quad k = \frac{(N+K)}{2}$$

Multiplication of $S_{yy}(f) \equiv S_k$ (or G_k) by the bandwidth of the signal $\Delta f = 1/(N+K)\Delta t$ gives the estimated signal variance, σ_k^2, in the kth frequency band; i.e. $\sigma_k^2 = S_k' = S_k \Delta f$. The summation

$$\sum_{n=0}^{N+K-1} S_k' = \sum_{n=0}^{N+K-1} S_k \Delta f \tag{5.6.19}$$

gives the variance and total power of the signal. The quantity

$$\begin{aligned} S_k' &= \frac{1}{[(N+K)\Delta t]^2}\left[|Y_k|^2 + |Y_{N+K-k}|^2\right] \\ &= \frac{1}{(N+K)^2}\sum_{n=0}^{N+K-1}\left|y_n e^{-i2\pi f n \Delta t}\right|^2 \end{aligned} \tag{5.6.20}$$

is often computed as the periodogram. However, this is not correctly scaled as a power spectral density but represents the "peak" in the spectral plot rather than the "area" under the plot of S_k versus Δf. The representation (5.6.20) is sometimes useful although most oceanographers are more familiar with the power spectral density form of the periodogram. It bears repeating that the use of Fourier transforms assumes a periodic structure to the sampled data when no periodic structure may actually exist in the time series. That is, the FFT of a finite length data record is equivalent to assuming that the record is periodic. We again note that autospectral functions are always real so that $S_{yy}'(f_k) = S_{yy}'(2f_N - f_k)$, and the one-sided autospectral periodogram estimate becomes

$$G_{yy}'(f_k) = 2S_k' = \frac{2}{[(N+K)\Delta t]^2}|Y(f_k)|^2 \tag{5.6.21}$$

Until the 1960s, the direct transform method first used by Schuster (1898) to study "hidden periodicities" in measured sun-spot numbers was seldom used due to difficulties with statistical reliability and extensive computational time. The introduction of the first practical FFT algorithms for spectral analysis (Cooley and Tukey, 1965) greatly reduced the computational time by taking advantage of patterns in discrete Fourier transform functions. Problems with the statistical reliability of the spectral estimates are resolved through appropriate windowing and averaging techniques which we discuss in Sections 5.6.6 and 5.6.7. Figure 5.6.8 compares the unsmoothed periodogram spectral estimate for the monthly mean sea surface temperature data at Amphitrite Point (Table 5.6.2) with the corresponding spectrum obtained from the Blackman–Tukey method. As mentioned earlier, the FFT requires data lengths equal to powers of 2 so that we have shortened the series to $2^5 = 32$

months. As we found with the Blackman–Tukey autocorrelation method, the FFT spectrum of coastal temperatures has a strong peak near the annual period, albeit with a slightly different spectral amplitude.

5.6.3.3 The power spectral density for periodic data

For a strictly periodic digital time series $y(t)$ having an exact integer number of oscillations over the interval $[0, T]$, we can use the Fourier series expansion (5.4.12) and write

$$y(t) = \frac{1}{2}A_0 + \sum_{n=1}^{N} [A_n \cos(\omega_n t) + B_n \sin(\omega_n t)] = \frac{1}{2}C_0 + \sum_{n=1}^{N} [C_n \cos(\omega_n t + \phi_n)] \quad (5.6.22)$$

in which the constants A_n, B_n are given by equation (5.4.14) and where

$$\begin{aligned} C_n &= (A_n^2 + B_n^2)^{1/2} \\ \phi_n &= \tan^{-1}(B_n/A_n) \end{aligned} \quad (5.6.23)$$

are the amplitude and phase of the complex Fourier coefficient for the nth frequency component, $\omega_n = 2\pi f_n$. Since the data record contains periodic components only, a plot of $2|C_n|^2$ against n ($n = 0, \ldots, N-1$) yields a series of distinct "spikes" or line spectra, S_n, with the variance divided equally between negative and positive frequencies

$$\begin{aligned} S_n &= \frac{(\Delta t)^2}{T}[|C_n|^2 + |C_{N-n}|^2] \\ &= \frac{2\Delta t}{N}|C_n|^2 \end{aligned} \quad (5.6.24)$$

where the record mean value C_0 has been subtracted from the record $y(t)$. Here we have assumed that $y(t)$ is a real function. The squared Fourier components $|C_n|^2$ give the contribution of the nth frequency component to the total variance and the various frequency components contribute additively to the total power of the time series. The contribution from each component is assumed to be independent of that from all other components.

5.6.3.4 Variance-preserving spectra

Because the power spectral density, $S_{yy}(f)$, of a time series often ranges over orders of magnitude, spectral distributions are usually plotted as the logarithm of $S_{yy}(f)$ versus frequency or the logarithm of frequency; i.e. $\log[S_{yy}(f)]$ versus f or $\log(f)$. The latter is especially useful where a spectrum has a power law dependence of the form $S_{yy}(f) \sim f^{-p}$. In this case, the slope of the spectrum is given as $p = -\log[S_{yy}(f)]/\log(f)$.

An example of $\log[S_{yy}(f)]$ versus f (a log–linear plot) is presented in Figure 5.6.9(a) where we have used time-series data generated by the relation $y(k) = A \cdot \cos(2\pi nk/N) + \varepsilon(k)$ from Section 5.6.1 (Figure 5.6.5a). The spectral density has units of energy/frequency for the same units used for f. For example, the PSD of a current velocity record are typically in units of $(\text{cm/s})^2/\text{cph}$ or $(\text{cm/s})^2/\text{cpd}$ plotted against frequency in cph (cycles per hour) or cpd (cycles per day), respectively. Sometimes, m/s are used in place of cm/s. Since the integration proceeds over frequency bands of width Δf centered at frequency f_c, the area under each small

rectangular segment of the spectral curve is equal to a pseudo-variance

$$\sigma_*^2(f_c) = \int_{f_c-\Delta f/2}^{f_c+\Delta f/2} \log\,[S_{yy}(f)]\,df \tag{5.6.25}$$

Although log spectra plots have an appealing shape, the integral (5.6.25) is certainly not variance-preserving. To preserve the signal variance, $\sigma^2(f_c)$, under the spectral curve, we need to plot $fS_{yy}(f)$ versus $\log(f)$ (Figure 5.6.9b). Replacing df in (5.6.25) with $d[\log(f)]$, the true *variance-preserving* form of the spectrum becomes

$$\sigma^2(f_c) = \int_{f_c-\Delta f/2}^{f_c+\Delta f/2} fS_{yy}(f)\,d[\log(f)] = \int_{f_c-\Delta f/2}^{f_c+\Delta f/2} S_{yy}(f)\,df \tag{5.6.26}$$

where we have used the fact that $d[\log(f)] = df/f$. Equation (5.6.26) gives the true signal variance within the band Δf. In particular, if $S_{yy}(f) \approx S_c$ is nearly constant over the frequency increment Δf, then $\sigma^2(f_c) \approx S_c \Delta f$ is the signal variance in band Δf centered at frequency f_c. In this format, there is a clear spectral peak at $f = 0.25$ cycles per unit time that is associated with the term $\cos(2\pi nk/N)$ in the original analytical expression.

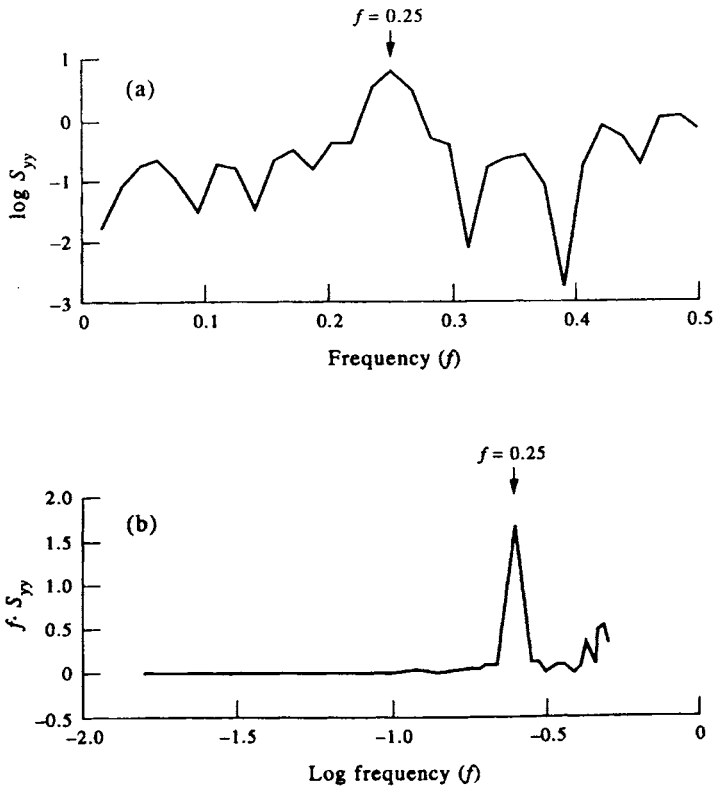


Figure 5.6.9. Two common types of spectral plot derived for the time series $y(k) = A\cos(2\pi nk/N) + \varepsilon(k)$ *(see Figure 5.6.5). (a) A plot of log power spectral density,* $\log[S_{yy}(f)]$*, versus frequency, f; (b) A variance-preserving plot,* $f \cdot [S_{yy}(f)]$ *versus* $\log(f)$.

5.6.3.5 The chi-squared property of spectral estimators

Throughout this chapter, we have claimed that each spectral estimate for maximum frequency resolution, $1/T$, obtained from Fourier transforms of stochastic time series have two degrees of freedom. We now present a more formal justification for that claim for discrete spectral estimators by showing that each estimate is a stochastic chi-square variable with two degrees of freedom (i.e. there are two independent squares entering the expression for the chi-square variable). Consider any stochastic white noise process $\eta(t)$, for which $E[\eta(t)] = 0$. The Fourier components are

$$\begin{aligned} A(f) &= \sum_{n=-N}^{N-1} \eta(n\Delta t) \cos(2\pi f n \Delta t) \\ B(f) &= \sum_{n=-N}^{N-1} \eta(n\Delta t) \sin(2\pi f n \Delta t) \end{aligned} \tag{5.6.27}$$

where as usual, $-1/(2\Delta t) \leq f \leq 1/(2\Delta t)$, and it follows that $E[A(f)] = 0 = E[B(f)]$. Thus, at the harmonic frequencies $f_k = k/N\Delta t$, the variance is

$$\begin{aligned} V[A(f_k)] = E[\mathrm{A}^2(f_k)] &= \sigma_\eta^2 \sum_{n=-N}^{N-1} \cos^2(2\pi f_k n \Delta t) \\ &= \tfrac{1}{2} N \sigma_\eta^2, \quad k = \pm 1, \pm 2, \ldots, \pm(N-1) \\ &= N\sigma_\eta^2, \quad k = 0, -N \end{aligned} \tag{5.6.28a}$$

Similarly

$$\begin{aligned} V[B(f_k)] &= \tfrac{1}{2} N \sigma_\eta^2, \quad k = \pm 1, \pm 2, \ldots, \pm(N-1) \\ &= 0, \quad k = 0, -N \end{aligned} \tag{5.6.28b}$$

When $k \neq j$, the covariance is

$$C[A(f_k), A(f_j)] = \sigma_\eta^2 \sum_{n=-N}^{N-1} \cos(2\pi f_k n \Delta t) \cos(2\pi f_j n \Delta t) = 0 \tag{5.6.29a}$$

and

$$C[A(f_k), B(f_j)] = 0 \text{ (orthogonality condition)} \tag{5.6.29b}$$

Because $A(f_k)$ and $B(f_k)$ are linear functions of normal random variables, $A(f_k)$ and $B(f_k)$ are also distributed normally. Hence, the random variables

$$\begin{aligned} \frac{A(f_k)^2}{V[A(f_k)]} &= \frac{2A(f_k)^2}{N\sigma_\eta^2} \\ \frac{B(f_k)^2}{V[B(f_k)]} &= \frac{2B(f_k)^2}{N\sigma_\eta^2} \end{aligned} \tag{5.6.30}$$

are each distributed as χ_1^2, which is a chi-square variable with one degree of freedom.

Since the normal distributions $A(f_k)$ and $B(f_k)$ are independent random variables, the sum of their squares

$$\frac{2}{\sigma_\eta^2}[A(f_k)^2 + B(f_k)^2] = \frac{2}{\Delta t \sigma_\eta^2} S_{yy}(f_k) \tag{5.6.31}$$

is distributed as χ_2^2, which is chi-square variable with two degrees of freedom. Here, $S_{yy}(f_k)$ is the sample spectrum. Thus

$$\frac{E[2S_{yy}(f_k)]}{\Delta t \sigma_\eta^2} = 2 \tag{5.6.32}$$

and

$$E[S_{yy}(f_k)] = \sigma_\eta^2 \Delta t \tag{5.6.33}$$

which is the spectrum. At the harmonic frequencies (set by the record length), the sample spectrum is an unbiased estimator of the white-noise spectrum of $\eta(t)$. Also, at these frequencies, the variance of the estimate is constant and independent of sample size. This explains the failure of the sample estimates of the variance to decrease with increasing sample size. We remark further that, even if $\eta(t)$ is not normally distributed, the random variables $A(f_k)$ and $B(f_k)$ are very nearly normally distributed by the central limit theorem. Hence, the distribution of the $S_{yy}(f)$ will be very nearly distributed as χ_2^2 regardless of the PDF of the $\eta(t)$ process.

5.6.4 Spectra of vector series

To calculate the spectra of vector time series such as current and wind, we first need to resolve the data into orthogonal components. Spectral analysis is then applied to the combined series of components and the results stored as a complex quantity in the computer. Raw data are recorded as speed and direction by rotor-type meters and as orthogonal components by acoustic and electromagnetic meters. The usual procedure is to convert recorded time series to an earth-referenced Cartesian coordinate system consisting of two orthogonal horizontal components and a vertical component. In the open ocean, horizontal velocities typically are resolved into components of eastward (zonal; u) and northward (meridional; v) time series, whereas in the coastal ocean it is preferable to resolve the vector components into cross-shore (u') and longshore (v') components through the rotation

$$\begin{pmatrix} u' \\ v' \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \tag{5.6.34a}$$

or

$$\begin{aligned} u' &= u\cos\theta + v\sin\theta \\ v' &= -u\sin\theta + v\cos\theta \end{aligned} \tag{5.6.34b}$$

where the angle θ is the orientation of the coastline (or the local bottom contours) measured counterclockwise from the eastward direction (Figure 5.6.10). Alternatively, one can let the current data define θ as the direction of the major axis obtained from principal component analysis; that is, the axis which maximizes the variance in a scatter plot of u versus v (see Figure 4.3.1).

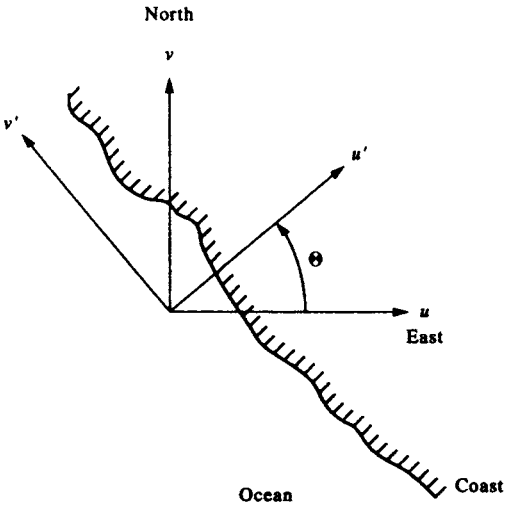


Figure 5.6.10. Cross-shore (u') and longshore (v') velocity components in a Cartesian coordinate system rotated through a positive (counterclockwise) angle from the eastward (u) and northward (v) directions.

In coastal regions, the principal axis is usually closely parallel to the coastline. For studies of highly circularly polarized motions, such as inertial waves and tidal currents, resolution into clockwise and counterclockwise rotary components is often more useful. The choice of representation depends on the preference of the investigator and the type of process being investigated. More will be said on this subject in Section 5.6.4.2.

5.6.4.1 Cartesian component rotary spectra

The horizontal velocity vector can be represented in Cartesian coordinates as a complex function $w(t)$ whose real part, $u(t)$, is the projection of the vector on the zonal (or cross-shelf) axis and whose imaginary part, $v(t)$, is the projection of the vector on the meridional (or longshelf) axis (Figure 5.6.11)

$$w(t) = u(t) + iv(t) \tag{5.6.35}$$

(The use of vector $w(t)$ follows the convention of Gonella (1972), Mooers (1973) and others in their discussion of rotary spectral analysis and is not to be confused with the weights $w(t)$ used in the sections on data windowing or the vertical component of velocity. Gonella (1972) used u_1 and u_2 for the two velocity components.) A complete description of the time variability of a three-dimensional vector at a single point consists of six functions of frequency: Three autospectra for the three velocity components and three cross-spectra. For the two-dimensional vectors considered in this section, there are two autospectra and one cross-spectrum. The discrete Fourier transform, $W(f_k) = U(f_k) + iV(f_k)$, ($f_k = k/N\Delta t$, $k = 1, \ldots, N$; $k = 0$ is the mean flow) is

$$\begin{aligned} W(f_k) &= \Delta t \sum_{n=0}^{N-1} w(t) e^{-i2\pi kn/N} \\ &= \Delta t \sum_{n=0}^{N-1} [u(t) + iv(t)] e^{-i2\pi kn/N} \end{aligned} \tag{5.6.36}$$

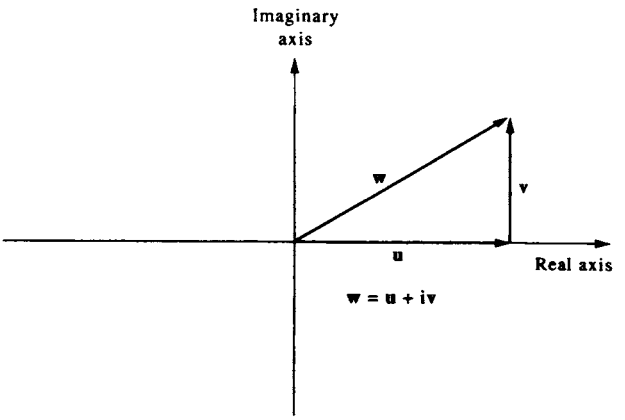


Figure 5.6.11. Horizontal velocity, w, represented as a complex vector $w = u + iv$ with components (u, v) along the real and imaginary axis, respectively.

where $U(f_k)$ and $V(f_k)$ are the Fourier transforms of $u(t)$ and $v(t)$, respectively. If the original record is separated into M blocks of length N', where $N = MN'$ is the total record length if no overlapping of segments is used, the spectral density function is given in terms of the number of segments used to form the block-averaged, one-sided autospectrum $(0 \leq f_k' < \infty)$

$$\begin{aligned} G_{ww}(f_k') &= \frac{2}{N\Delta t}\sum_{m=1}^{M}|W_m(f_k')|^2 \\ &= \frac{2}{N\Delta t}\sum_{m=1}^{M}\left\{[W_{Rm}(f_k')]^2 + [W_{Im}(f_k')]^2\right\} \\ &= \frac{2}{N\Delta t}\sum_{m=1}^{M}\left\{[U_{Rm}(f_k') - V_{Im}(f_k')]^2 + [U_{Im}(f_k') + V_{Rm}(f_k')]^2\right\} \end{aligned} \tag{5.6.37}$$

where $f_k' = k/N'\Delta t$, $k = 0, 1, \ldots, N'/2$ ($k = 0$ is the mean flow) and for FFT analysis, $N' = 2p$ (positive integer p), and where the subscripts R and I stand for the real and imaginary parts of the given Fourier components.

5.6.4.2 Rotary component spectra

Rotary analysis of currents involves the separation of the velocity vector for a specified frequency, ω, into clockwise and counterclockwise rotating circular components with amplitudes A^-, A^+ and relative phases θ^-, θ^+, respectively. Thus, instead of dealing with two Cartesian components (u, v) we deal with two circular components $(A^-, \theta^-; A^+, \theta^+)$. Several reasons can be given for using this approach: (1) the separation of a velocity vector into oppositely rotating components can reveal important aspects of the wave field at the specified frequencies. The method has proven especially useful for investigating currents over abrupt topography, wind-generated inertial motions, diurnal frequency continental shelf waves, and other forms of narrow-band oscillatory flow; (2) in many cases, one of the rotary components (typically, the clockwise component in the northern hemisphere and counterclockwise

component in the southern hemisphere) dominates the currents so that we need only deal with one scalar quantity rather than two. Inertial motions, for example, are almost entirely clockwise rotary in the northern hemisphere so that the counterclockwise component can be ignored for most applications; (3) many of the rotary properties, such as spectral energy $S^-(\omega)$ and $S^+(\omega)$ and rotary coefficient, $r(\omega)$, are invariant under coordinate rotation so that local steering of the currents by bottom topography or the coastline are not factors in the analysis.

The vector addition of the two oppositely rotating circular vectors (Figure 5.6.12a, b) causes the tip of the combined vector (Figure 5.6.12c) to trace out an ellipse over one complete cycle. The eccentricity, e, of the ellipse is determined by the relative amplitudes of the two components. Motions at frequency ω are circularly polarized if one of the two components is zero; motions are rectilinear (back-and-forth along the same line) if both circularly polarized components have the same magnitude. In rotary spectral format, the current vector $w(t)$ can be written as the Fourier series

$$\begin{aligned} w(t) &= \overline{u(t)} + \sum_{k=1}^{N} U_k \cos(\omega_k t - \phi_k) + \mathrm{i}\left[\overline{v(t)} + \sum_{k=1}^{N} V_k \cos(\omega_k t - \theta_k)\right] \\ &= [\overline{u(t)} + \mathrm{i}\overline{v(t)}] + \sum_{k=1}^{N} [U_k \cos(\omega_k t - \phi_k) + \mathrm{i} V_k \cos(\omega_k t - \theta_k)] \end{aligned} \tag{5.6.38}$$

in which $\overline{u(t)} + \mathrm{i}\overline{v(t)}$is the mean velocity, $\omega_k = 2\pi f_k = 2\pi k/N\Delta t$ is the angular frequency, t $(= n\Delta t)$ is the time and (U_k, V_k) and (ϕ_k, θ_k) are the amplitudes and phases, respectively, of the Fourier constituents for each frequency for the real and imaginary components. Subtracting the mean velocity and expanding the trigonometric functions, we find

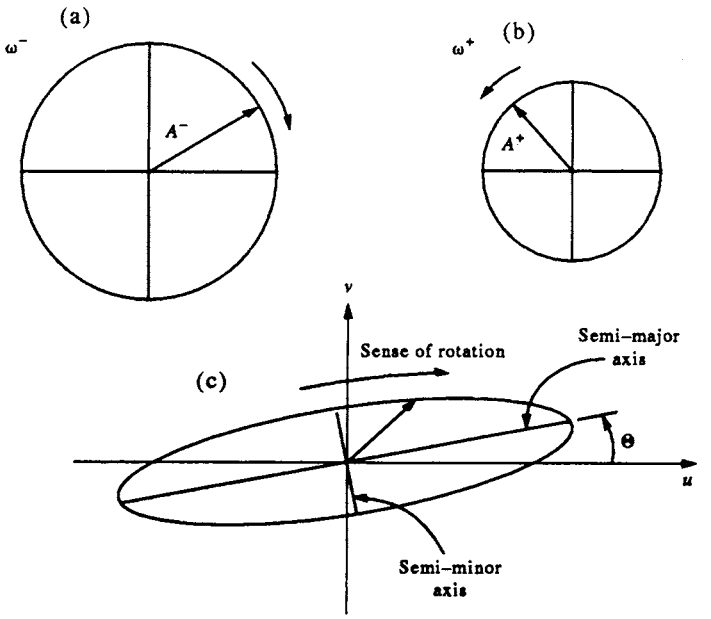


Figure 5.6.12. Current ellipses formed by the vector addition of two oppositely rotating vectors. (a) Clockwise component (ω^-) *and (b) counterclockwise component* (ω^+) *with amplitudes* A^- *and* A^+*, respectively. (c) General case of elliptical motion with major axis tilted at an angle* θ *counterclockwise from east.* ε^- *and* ε^+ *(not shown) are the angles of the two circular components at time* $t = 0$*.*

$$w'(t) = w(t) - \left[\overline{u(t)} + i\overline{v(t)}\right]$$
$$= \sum_{k=1}^{N} \{U_{1k}\cos(\omega_k t) + U_{2k}\sin(\omega_k t) + i[V_{1k}\cos(\omega_k t) + V_{2k}\sin(\omega_k t)]\} \tag{5.6.39}$$

in which we have defined the even (U_{1k}, V_{1k}) and odd (U_{2k}, V_{2k}) functions as

$$U_{1k} = U_k\cos\phi_k, \quad U_{2k} = U_k\sin\phi_k \tag{5.6.40a}$$

$$V_{1k} = V_k\cos\theta_k, \quad V_{2k} = V_k\sin\theta_k \tag{5.6.40b}$$

Dropping the prime notation for $w'(t)$ and following some reorganization, we can write the kth frequency component of the series as the sum of counterclockwise (+) and clockwise (−) components

$$\begin{aligned} w_k(t) &= w_k^+(t) + w_k^-(t) \\ &= A_k^+\exp(i\varepsilon_k^+)\exp(i\omega_k t) + A_k^-\exp(i\varepsilon_k^-)\exp(-i\omega_k t) \\ &= \exp\left[\frac{i(\varepsilon_k^+ + \varepsilon_k^-)}{2}\right]\left\{[A_k^+ + A_k]\cos\left[\frac{\varepsilon_k^+ - \varepsilon_k^-}{2} + \omega_k t\right]\right. \\ &\left. + i[A_k^+ - A_k^-]\sin\left[\frac{\varepsilon_k^+ - \varepsilon_k^-}{2} + \omega_k t\right]\right\} \end{aligned} \tag{5.6.41}$$

where the counterclockwise and clockwise rotary component amplitudes are given by

$$A_k^+ = \frac{1}{2}\left\{[(U_{1k} + V_{2k})]^2 + [(U_{2k} - V_{1k})]^2\right\}^{1/2} \tag{5.6.42a}$$

$$A_k^- = \frac{1}{2}\left\{[(U_{1k} - V_{2k})]^2 + [(U_{2k} + V_{1k})]^2\right\}^{1/2} \tag{5.6.42b}$$

and the corresponding phase angles for time $t = 0$, by

$$\varepsilon_k^+ = \tan^{-1}[(V_{1k} - U_{2k})/(U_{1k} + V_{2k})] \tag{5.6.43a}$$

$$\varepsilon_k^- = \tan^{-1}[(U_{2k} + V_{1k})/(U_{1k} - V_{2k})] \tag{5.6.43b}$$

Each of the constituents contributing to equation (5.6.39) have the form of an ellipse with major semi-axis of length $L_M = (A_k^+ + A_k^-)$ and minor semi-axis of length $L_m = |A_k^+ - A_k^-|$ (Figure 5.6.12c). The ellipse is tilted at an angle of $\theta = \frac{1}{2}(\varepsilon_k^+ + \varepsilon_k^-)$ from the u-axis and the vector is along the major axis of the ellipse at time $t = (\varepsilon_k^+ - \varepsilon_k^-)/(4\pi f_k)$. The one-sided spectra $(G_k^+, G_k^-) = (S_k^+, S_k^-)$ for the two oppositely rotating components for frequencies $f_k = \omega_k/2\pi$ are

$$S(f_k^+) = S_k^+ = \frac{(A_k^+)^2}{N\Delta t}, \quad f_k = 0, \ldots, 1/(2\Delta t) \tag{5.6.44a}$$

$$S(f_k^-) = S_k^- = \frac{(A_k^-)^2}{N\Delta t}, \quad f_k = -1/(2\Delta t), \ldots, 0 \tag{5.6.44b}$$

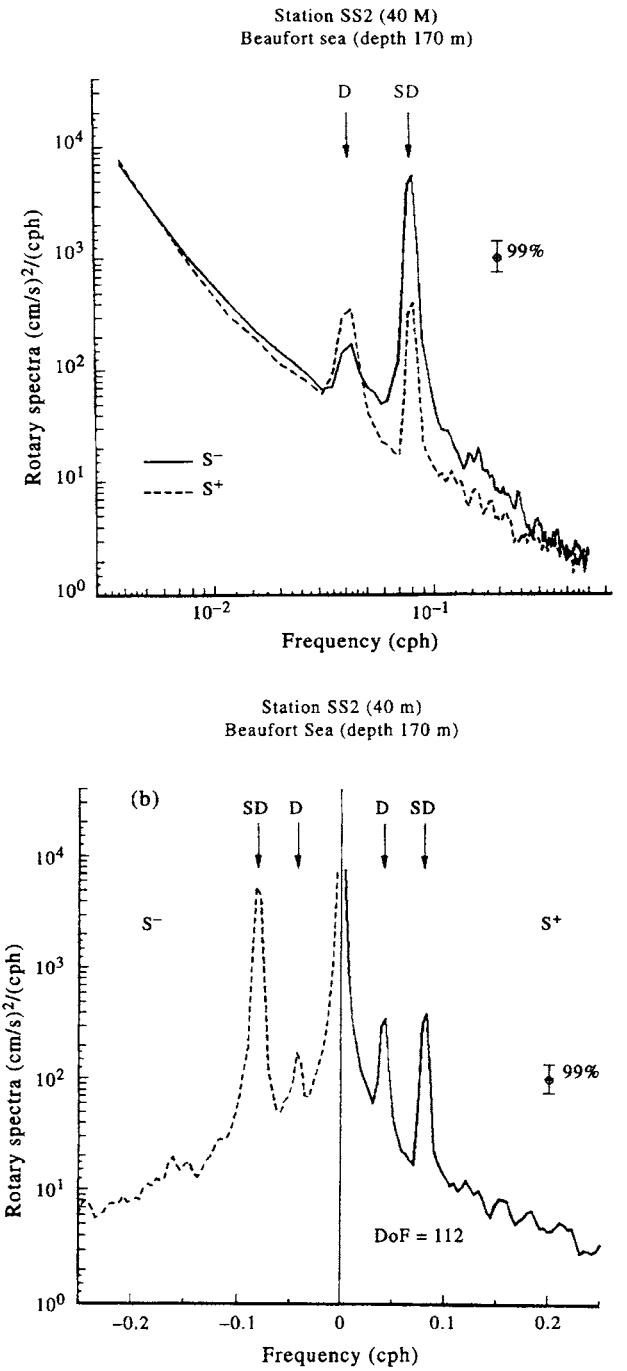


Figure 5.6.13. Rotary current spectra for hourly currents measured at 40-m depth in the Beaufort Sea, Arctic Ocean (water depth = 170 m). Peaks are at the diurnal (D) and semidiurnal (SD) tidal frequencies. Frequency resolution is 0.0005 cph and there are 112 degrees of freedom per spectral band. Vertical bar gives the 99% level of confidence. (a) One-sided rotary spectra, $S^-(f)$ and $S^+(f)$, versus f for positive frequency, f; (b) two-sided rotary spectra, $S(f_k^+)=S^+$ and $S(f_k^-)=S^-$ versus log$f$ for positive and negative frequencies, $f_k^{\pm}$. (Courtesy E. Carmack, A. Rabinovich, and E. Kolikov.)

Plots of rotary spectra are generally presented in two ways. In Figure 5.6.13(a), both S^- and S^+ are plotted as functions of frequency magnitude, $|f| \geq 0$, with solid and dashed lines used for the clockwise and counterclockwise spectra, respectively. In Figure 5.6.13(b), we use the fact that clockwise spectra are defined for negative frequencies and counterclockwise spectra for positive frequencies. The spectra $S(f_k^+)$ and $S(f_k^-)$ used in Figure 5.6.13(a) are then plotted on opposite sides of zero frequency. In these spectra, peak energy occurs at the diurnal and semidiurnal periods. The predominantly clockwise rotary motions at semidiurnal periods suggest a combination of tidal and near-inertial motions (at this latitude the inertial period is close to the semidiurnal tidal period).

Another useful property is the rotary coefficient

$$r(\omega) = \frac{S_k^+ - S_k^-}{S_k^+ + S_k^-} \tag{5.6.45}$$

which ranges from $r = -1$ for clockwise motion, to $r = 0$ for unidirectional flow, to $r = +1$ for counterclockwise motion. The rotary nature of the flow can change considerably with position, depth and time. As indicated by Figure 5.6.14, the observed diurnal tidal currents over Endeavour Ridge in the northeast Pacific change from moderately positive to strongly negative rotation with depth. In contrast, the semidiurnal currents change from strongly negative near the surface to strongly rectilinear at depth. (Data, in this case, are from a string of current meters moored for a period of nine months in the northeast Pacific.) We remark that the definition (5.6.45) differs in sign from that of Gonella (1972) who used $S_k^- - S_k^+$ rather than $S_k^+ - S_k^-$ in the numerator. Because many types of oceanic flow are predominantly clockwise rotary in the northern hemisphere, Gonella's definition has the advantage that clockwise rotating currents have positive rotary coefficients. However, we find Gonella's definition confusing since clockwise motions, which are linked to negative frequencies, then have positive rotary coefficients.

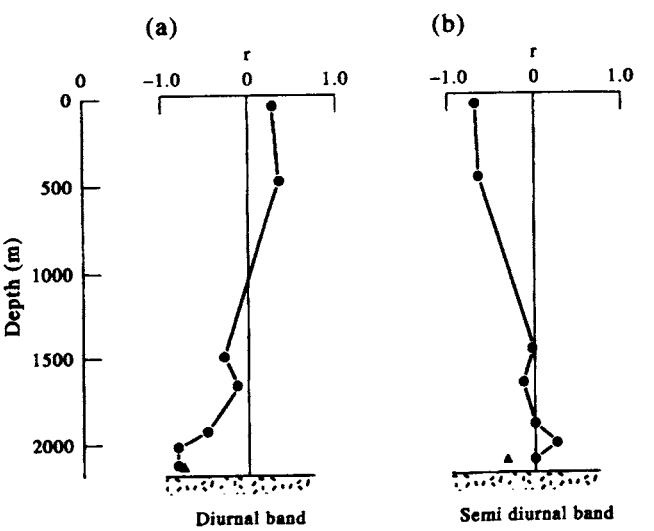


Figure 5.6.14. Rotary coefficient, $r(\omega)$, as a function of depth for current oscillations in (a) the diurnal frequency band ($\omega/2\pi \approx 0.04$ cph) and (b) the semidiurnal band ($\omega/2\pi \approx 0.08$ cph). (From Allen and Thomson, 1993.)

5.6.4.3 Rotary spectra (via Cartesian components)

Gonella (1972) and Mooers (1973) present the rotary spectra in terms of their Cartesian counterparts and provide a number of rotational invariants for analyzing current and wind vectors at specified frequencies. Specifically, the one-side auto-spectra for the counterclockwise (CCW) and clockwise (CW) rotary components of the vector $w(t) = u(t) + iv(t)$ are, in terms of their Cartesian components

$$G(f_k^+) = \tfrac{1}{2}[G_{uu}(f_k) + G_{vv}(f_k) + Q_{uv}(f_k)], \quad f_k \geq 0 \text{ (CCW component)} \tag{5.6.46a}$$

$$G(f_k^-) = \tfrac{1}{2}[G_{uu}(f_k) + G_{vv}(f_k) - Q_{uv}(f_k)], \quad f_k \leq 0 \text{ (CW component)} \tag{5.6.46b}$$

where $G_{uu}(f_k)$ and $G_{vv}(f_k)$ are the one-sided autospectra of the u and v Cartesian components of velocity and $Q_{uv}(f_k)$ is the quadrature spectrum between the two components, where

$$Q_{uv}(f_k) = -Q_{uv}(-f_k) = (U_{1k}V_{2k} - V_{1k}U_{2k}) \tag{5.6.47}$$

As defined in Section 5.8, the spectrum can be written in terms of co-spectrum (real part) and quadrature spectrum (imaginary part)

$$G_{uv}(f_k) = C_{uv}(f_k) - iQ_{uv}(f_k) \tag{5.6.48}$$

5.6.5 Effect of sampling on spectral estimates

Spectral estimates derived by conventional techniques are limited by two fundamental problems: (1) the finite length, T, of the time series; and (2) the discretization using the sampling interval, Δt. The first problem is inherent to all real datasets while the second is associated with finite instrument response times and/or the need to digitize the time series for purposes of analysis.

Irrespective of the method used to calculate the power spectrum of a waveform, the record duration $T = N\Delta t$ and sampling increment Δt impose severe limitations on the information that can be extracted. Ideally, we would like to sample rapidly enough (small Δt) that no significant frequency component goes unresolved. This also eliminates aliasing problems in which unresolved spectral energy at frequencies higher than the Nyquist frequency is folded back into lower frequencies. At the same time we wish to record for a sufficiently long period (large N) that we capture many cycles of the lowest frequency of interest. Long-term sampling also enables us to better resolve frequencies that are close together and to improve the statistics (confidence intervals) for spectral estimates. In reality, most data series are a compromise based on the frequencies of interest, the response limitations of the sensor, and cost. The choices of the sampling rate and the record duration are tailored to best meet the task at hand.

5.6.5.1 Effect of finite record length

As noted earlier, we can think of a data sample $\{y(t)\}$ of duration $T = N\Delta t$ as the output from an infinite physical process $\{y'(t)\}$ viewed through a finite length window (Figure 5.6.1). The window has the shape of a "box-car" function $w(t_n) = w_n =$

$w(n\Delta t)$ which has unit amplitude and zero phase lag over the duration of the data sequence but is zero elsewhere. That is $y(t_n) = w(t_n)y'(t_n)$ where

$$\begin{aligned} w_n &= 1.0, \quad n = 0, \ldots, N-1 \\ w_n &= 0, \quad \text{for } n \geq N, n < 0 \end{aligned} \tag{5.6.49}$$

Since it is truncated, the dataset has endpoint discontinuities which lead to Gibbs' phenomena "ringing" and the ripple effects in the frequency domain. The discrete Fourier transform $Y(f)$ of the truncated series $y_n = y(n\Delta t)$ is

$$Y(f) = \sum_{n=-\infty}^{\infty} w_n y'_n e^{-i2\pi f n \Delta t} \tag{5.6.50}$$

In frequency space, $Y(f)$ is the convolution (written as $*$) of the Fourier transform of the infinite data set, $Y'(f)$, with the Fourier transform $W(f)$ of the function $w(t)$. That is

$$\begin{aligned} Y(f) &= \int_{-\infty}^{\infty} Y'(f')W'(f-f')\,df \\ &= Y'(f) * W(f) \end{aligned} \tag{5.6.51}$$

where for a box-car function

$$\begin{aligned} W(f) &= T \exp(i\pi fT) \frac{\sin(\pi f N \Delta t)}{(\pi f N \Delta t)} \\ &= T \exp(i\pi fT) \operatorname{sinc}(\pi f N \Delta t) \end{aligned} \tag{5.6.52}$$

and $\operatorname{sinc}(x) \equiv \sin(x)/x$. It is the large side-lobes or ripples of the sinc function (Figure 5.6.15) which are responsible for the leakage of spectral power from the main frequency components into neighboring frequency bands (and vice versa). In particular, $Y(f)$ for a specific frequency $f = f_o$ is spread to other frequencies, f, according to the phase and amplitude weighting of the window function. Leakage has the effect of both reducing the spectral power in the central frequency component and contaminating it with spectral energy from adjacent frequency bands. Those familiar with the various mathematical forms for the Dirac delta function, $\delta(f)$, will recognize the formulation

$$\delta(f) = \lim_{f \to 0} \left[\frac{\sin(\pi f \Delta t)}{\pi f \Delta t} \right]$$

Thus, as the frequency resolution increases (i.e. $f \to 0$), $Y(f) \to Y'(f)$.

In addition to distorting the spectrum, the box-car window limits the frequency resolution of the periodogram, independently of the data. The convolution $Y'(f) * W(f)$ means that the narrowest spectral response of the resultant transform is confined to the main-lobe width of the window transform. For a given window, the main-lobe width (the width between the -3 dB levels of the main lobe) determines the frequency resolution, Δf, of a particular window. For most windows, including the box-car window, this resolution is roughly the inverse of the observation time; $\Delta f \approx 1/T = 1/N\Delta t$.

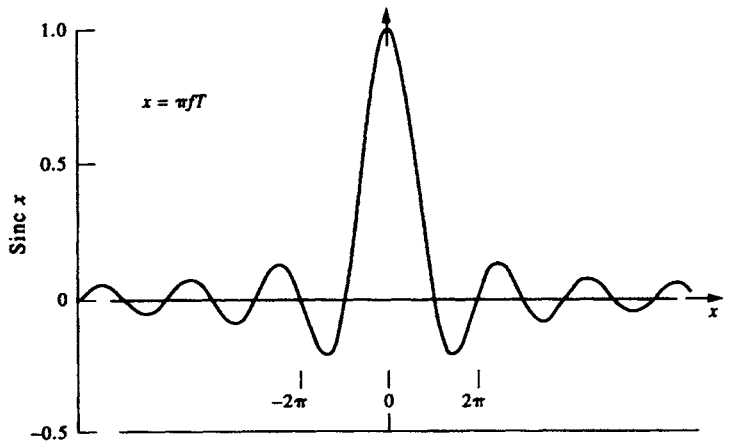


Figure 5.6.15. The function sinc (x) = sin (x)/x showing the large side-lobes which are responsible for leakage of spectral power from a given frequency to adjacent frequencies.

5.6.5.2 Aliasing

Poor discretization of time-series data due to limitations in the response time of the sensor, limitations in the recording and data storage rates, or through post-processing methods may cause *aliasing* of certain frequency components in the original waveform (Figure 5.6.16a). An aliased frequency is one that masquerades as another frequency. In Figure 5.6.16(b), for example, the considerable tidal energy at diurnal and semidiurnal periods (1 and 2 cpd) is folded back to lower frequencies of 0.07 and 0.10 cpd that are nowhere near the original frequencies. For a specific sampling interval, it becomes impossible to tell with certainty which frequency out of a large number of possible aliases is actually contributing to the signal variability. This leads to differences in the spectra between the continuous and discrete time series. Since we use the spectra of the discrete series to estimate the spectrum of the continuous series, the sampling interval must be properly selected to minimize the effect of the aliasing. If we know from previous analysis that there is little likelihood of significant energy at the disguised frequencies, then aliasing is not a problem. Otherwise, a degree of smoothing may be required to ensure that higher frequencies do not contaminate the lower frequencies. This smoothing must be performed prior to sampling or digitizing since aliased contributions cannot be recognized once they are present in the discrete data series.

The aliasing problem can be illustrated in a number of ways. To begin with, we note that for discrete data at equally spaced intervals Δt, we can measure only those frequency components lying within the principal frequency range,

$$-\omega_N \leq \omega \leq -\omega_o, \quad \omega_o \leq \omega \leq \omega_N, \quad \omega_N \geq 0 \tag{5.6.53a}$$

$$-f_N \leq f \leq -f_o, \quad f_o \leq f \leq f_N, \quad f_N \geq 0 \tag{5.6.53b}$$

in which $\omega_N = \pi/\Delta t$ and $f_N = 1/(2\Delta t)$ are the usual Nyquist frequencies in radians and cycles per unit time, respectively, and $\omega_o = 2\pi/T$ and $f_o = 1/T$ are corresponding fundamental frequencies for a time series of duration T. The Nyquist frequency is the highest frequency that can be extracted from a time series having a sampling rate of $1/\Delta t$. Clearly, if the original time series has spectral power at frequencies for which

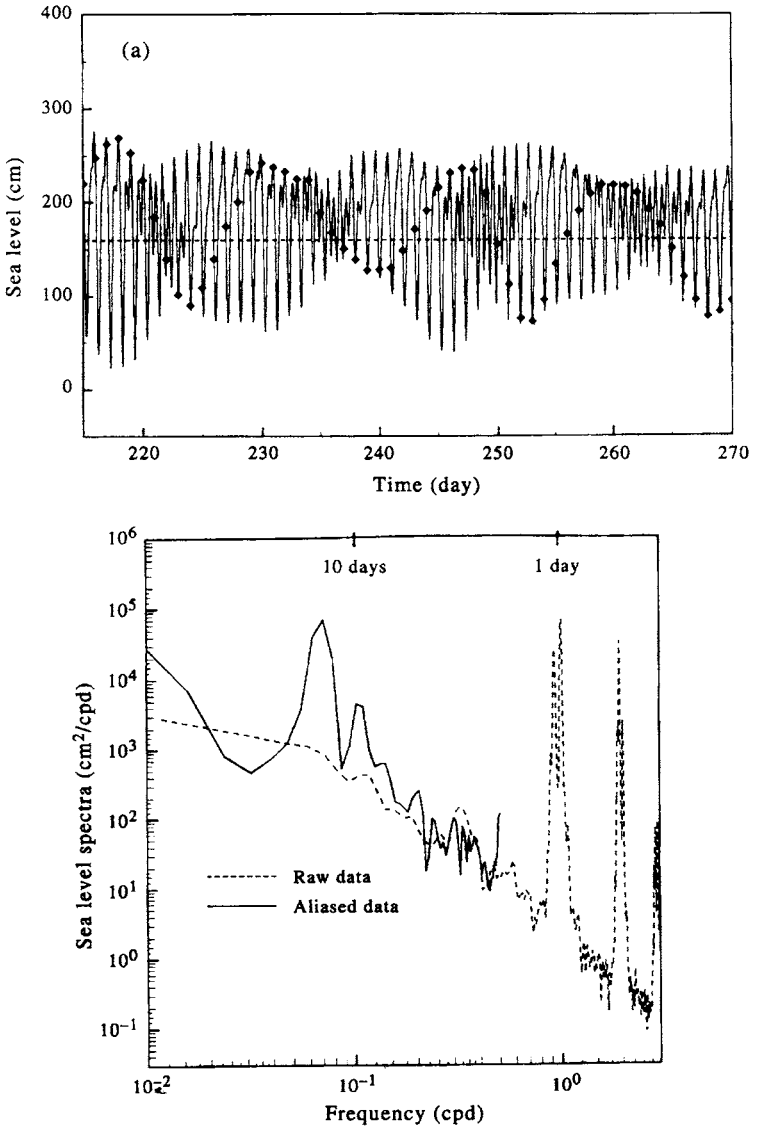


Figure 5.6.16. The origin of alaising. (a) The solid line is the tide height recorded at Victoria, British Columbia over a 60-day period from 29 July to 27 September 1975 (time in Julian days). The diamonds are the sea-level values one would obtain by only sampling once per day.(b) The power spectrum obtained from the two data series in (a). In this case, the high frequency energy (dashed curve) gets folded back into the spectrum at lower (aliased) frequencies (solid curve).

$|f| \geq f_N$, these spectral contributions are unresolved and will contaminate power associated with frequencies within the principal range (Figure 5.6.17). The unresolved variance becomes lumped together with other frequency components. Familiar examples of aliasing are the slow reverse rotation of stage-coach wheels in classic western movies due to the under-sampling by the frame-rate of the movie camera. Even in modern film, distinguishable features on moving automobile tires often can be seen to rotate rapidly backwards, slow to a stop, then turn forward at the correct

rotation speed as the vehicle gradually comes to a stop. Automobile commercials avoid this problem by equipping the wheels with featureless hubcaps.

If $\omega, f \geq 0$ are frequencies inside the principal intervals (5.6.53), the frequencies outside the interval which form aliases with these frequencies are

$$2\omega_N \pm \omega, \quad 4\omega_N \pm \omega, \ ..., \ 2p\omega_N \pm \omega \tag{5.6.54a}$$

$$2f_N \pm f, \quad 4f_N \pm f, \ ..., \ 2pf_N \pm f \tag{5.6.54b}$$

where p is a positive integer. These results lead to the alternate term *folding* frequency for the Nyquist frequency since spectral power outside the principal range is folded back, accordion-style, into the principal interval. As illustrated by Figure 5.6.17, folding the power spectrum about f_N produces aliasing of frequencies $2f_N - f$ with frequencies f; folding the spectrum at $2f_N$ produces aliasing of frequencies $2f_N + f$ with frequencies $2f_N - f$ which are then folded back about f_N into frequency f, and so forth. For example, if $f_N = 5$ rad/h, the observations at 2 rad/h are aliased with spectral contributions having frequencies of 8 and 12 rad/h, 18 and 22 rad/h, and so on.

We can verify that oscillations of frequency $2p\omega_N \pm \omega$ (or $2pf_N \pm f$) are indistinguishable from frequency ω (or f) by considering the data series $x_\omega(t)$ created by the single frequency component $x_\omega(t) = \cos(\omega t)$. Using the transformation $\omega \rightarrow (2p\omega_N \pm \omega)$, together with $t_n = n\Delta t$ and $\omega_N = \pi/\Delta t$, yields

$$\begin{aligned} x_\omega(t_n) &= \cos[(2p\omega_N \pm \omega)t_n] = \text{Re}\{\exp[\text{i}(2p\omega_N \pm \omega)t_n]\} \\ &= \text{Re}\{\exp[\text{i}2p\omega_N t_n]\exp[\pm\text{i}\omega t_n]\} \\ &= (+1)^{pn}\text{Re}[\exp(\pm\text{i}\omega t_n)] = \cos(\omega t_n) = x_\omega(t_n) \end{aligned} \tag{5.6.55}$$

In other words, the spectrum of $x(t)$ at frequency ω will be a superposition of spectral contributions from frequencies ω, $2p\omega_N \pm \omega$, $4p\omega_N \pm \omega$, and so forth. More specifically, it can be shown that the aliased spectrum $S_a(\omega)$ for discrete data is given by

$$S_a(\omega) = \sum_{n=-\infty}^{\infty} S(\omega + 2n\omega_N) \tag{5.6.56a}$$

$$= S(\omega) + \sum_{n=1}^{\infty} [S(2n\omega_N - \omega) + S(2n\omega_N + \omega)] \tag{5.6.56b}$$

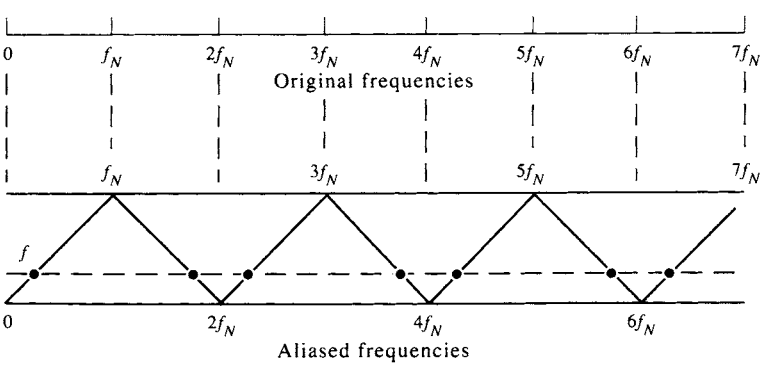


Figure 5.6.17. The spectral energies of all frequencies $f = \omega/2\pi$ at the nodes (•) located along the dotted line are folded back, accordion style, into the spectral estimate for the spectrum $S(f)$ for the primary range $0 \leq f \leq f_N$ ($0 \leq \omega \leq \omega_N$). (Adapted from Bendat and Piersol, 1986.)

The true spectrum, S, gives the distorted spectrum, S_a, caused by the summation of overlapping copies of measured spectra in the principal interval. Only if the original record is devoid of spectral power at frequencies outside the principal frequency range will the spectrum of the observed record equal that of the actual oceanic variability. To avoid aliasing problems, one has no choice but to sample the data as frequently as justifiably possible (i.e. up to frequencies beyond which energy levels become small) or to filter the sampled data before they are recorded (as in the case of a stilling well used to eliminate gravity waves from a tidal record). A further example of spectral contamination by aliased frequencies is illustrated in Figure 5.6.18(a, b). In Figure 5.6.18(b), we have assumed that the wave recorder was inadvertently programmed to record at 0.15 Hz, corresponding to a limiting wave period of 6.67 s. The energy from the shorter period waves were not measured but contaminate the energy of the longer period waves when folded back about the Nyquist frequency.

5.6.5.3 Nyquist frequency sampling

Sampling time series that have significant variability at the Nyquist frequency affords its own set of problems. Suppose we wish to represent $y(t)$ through the usual Fourier relation

$$y(t) = \int_{-\omega_N}^{\omega_N} Y(\omega) e^{i\omega t}\, d\omega \tag{5.6.57}$$

where we have assumed that $Y(\omega) = 0$ for $|\omega| > \omega_N$. In this case, there is no aliasing problem since there is no power at frequencies greater than ω_N. The function $y(t)$ can be constructed from frequency components strictly in the interval $(-\omega_N, \omega_N)$. In discrete form for infinite length data

$$y(t) = \frac{1}{2\omega_N} \sum_{n=-\infty}^{\infty} \left[y_n \int_{-\omega_N}^{\omega_N} e^{i\omega(t-n\Delta t)}\, d\omega \right] \tag{5.6.58a}$$

where the integral has the form of a sinc function such that

$$y(t) = \sum_{n=-\infty}^{\infty} y_n \frac{\sin[\omega_N(t - n\Delta t)]}{t - n\Delta t} \tag{5.6.58b}$$

Given the data $\{y_n\}$, we can construct $y(t)$. However, suppose that $y(t)$ fluctuates with the Nyquist frequency ω_N such that

$$y(t) = y_o \cos(\omega_N t + \theta) \tag{5.6.59}$$

where, for the sake of generality, the phase angle is arbitrary, $0 \leq \theta \leq 2\pi$. Then, using $\sin(n\pi) = 0$ for all n (an integer)

$$\begin{aligned} y_n &= y(n\Delta t) = y_o \cos(n\pi + \theta) = y_o[\cos(n\pi)\cos\theta] \\ &= y_o(-1)^n \cos\theta \end{aligned} \tag{5.6.60}$$

This leads to a component with amplitude $y_n = y_o(-1)^n \cos\theta$ which fluctuates in sign because of the term $(-1)^n$, $-\infty \leq n \leq \infty$. If θ is unknown, the function $y(t)$ cannot be

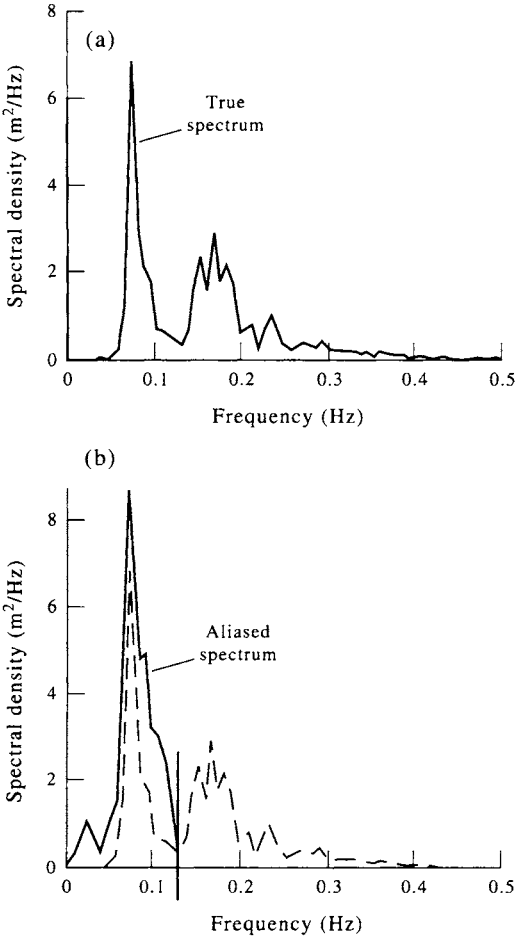


Figure 5.6.18. An aliased autospectrum. (a) The true spectrum, S(f) (m^2/cps), of wind-generated waves as a function of frequency (Hz = cycles per second); (b) Aliased spectrum, $S_a(f)$, that would arise from folding about a hypothetical Nyquist frequency f_N = 0.13 Hz.

constructed. If $\theta = k\pi/2$, so that $\cos(\omega_N t + \theta) = \sin(\omega_N t)$, the observer will find no signal at all. In general, $0 \leq |\cos\theta| < 1$ and the magnitude will always be less than y_o, resulting in biased data.

According to the above analysis, we should sample slightly more frequently than Δt if we are to fully resolve oscillations at the maximum frequency of interest (assumed to be the Nyquist frequency). A sampling rate of 2.5 samples per cycle of the frequency of interest appears to be acceptable whereby $\Delta t = 1/(2.5f_N) = (2/5)(1/f_N) = (4/5)\pi/\omega_N$.

5.6.5.4 Frequency resolution

The need to resolve spectral estimates in neighboring frequency bands is an important requirement of time series analysis. Without sufficient resolution, it is not possible to determine whether a given spectral peak is associated with a single frequency, or is a smeared response containing a number of separate spectral peaks. A good example of this for tides is presented by Munk and Cartwright (1966) who show that for long records the main constituents in the diurnal and semidiurnal frequency bands can be

resolved into a multitude of other tidal frequencies. How well the peaks can be resolved depends on the frequency differences, Δf, between the peaks and the length, T, of the data set used in the analysis. For an unsmoothed periodogram, the frequency resolution in hertz is roughly the reciprocal of the time duration in seconds of the data.

The distinction between well-resolved and poorly resolved spectral estimates is somewhat subjective and depends on how we wish to define "resolution". As with diffraction patterns in classical optics, we can follow the "Rayleigh criterion" for the separation of spectral peaks (Jenkins and White, 1957). Recall that the diffraction pattern for a given frequency, f, of light varies as $\text{sinc}(\phi) = \sin[(\phi - \phi_f)]/(\phi - \phi_f)$, where ϕ is the angle of the incident light beam to the grating. This also is the functional form for the spectral peak of a truncated time series (see *windowing* in the next section). Two spectral lines are said to be "well resolved" if the separation between peaks exceeds the difference in frequency between the center frequency to the maximum at the first side-lobe and "just resolved" if the spectral peak of one pattern coincides with the first zero of the second pattern (Figure 5.6.19a–c). Here, the separation in frequency is equal to the difference in frequency between the peak of one spectrum and the first zero of the function $\sin(\phi)/\phi$ of the second (where $\phi = \omega T/2$). The spectral peaks are "not resolved" if this separation is less than the separation between the center frequency and the first zero of the $\sin(\phi)/\phi$ functions (Figure 5.6.19d).

Consider an oceanic record consisting of two sinusoidal components, both having amplitude y_o and constant phase lags such that

$$y(t) = y_o[\cos(\omega_1 t + \theta_1) + \cos(\omega_2 t + \theta_2)], \quad -T/2 \le t \le T/2 \tag{5.6.61}$$

where as usual $\omega = 2\pi f$. The one-sided, unsmoothed power spectral density, $S(\omega)$, for these data are then found from the Fourier transform

$$S(\omega) = \tfrac{1}{2} T y_o^2 \left\{ \frac{\sin[\frac{1}{2}T(\omega - \omega_1)]}{[\frac{1}{2}T(\omega - \omega_1)]} + \frac{\sin[\frac{1}{2}T(\omega - \omega_2)]}{[\frac{1}{2}T(\omega - \omega_2)]} \right\}$$

The power spectrum consists of two terms of the form $\sin(\phi)/\phi$ centered at frequencies ω_1 and ω_2. Using the Rayleigh criterion, we can just resolve the two peaks (i.e. determine if there is one or two sinusoids contributing to the spectrum) provided that the frequency separation $\Delta\omega = |\omega_1 - \omega_2|$ ($\Delta f = |f_1 - f_2|$) is equal to the frequency difference for the peak of one frequency and the first zero of $\sin(\phi)/\phi$ for the other frequency. Since zeros of $\sin(\phi)/\phi$ occur at frequencies f equal to $\pm 1/T$, $\pm 2/T$, ..., $\pm p/T$, the frequencies are just resolved when

$$\Delta\omega = \frac{2\pi}{T}; \quad \Delta f = \frac{1}{T} \tag{5.6.63a}$$

and well-resolved for

$$\Delta\omega > \frac{3\pi}{T}; \quad \Delta f > \frac{3}{2T} \tag{5.6.63b}$$

In summary, resolution of two frequencies f_k and $f_{k+1} (= f_k \pm \Delta f)$ using an unsmoothed periodogram or equivalently a rectangular window, requires a record of length T, where $\Delta f = 1/T$ frequency units. Note also that $1/T$ is equal to the fundamental frequency, f_1, which is the lowest frequency that we can calculate for the record. For

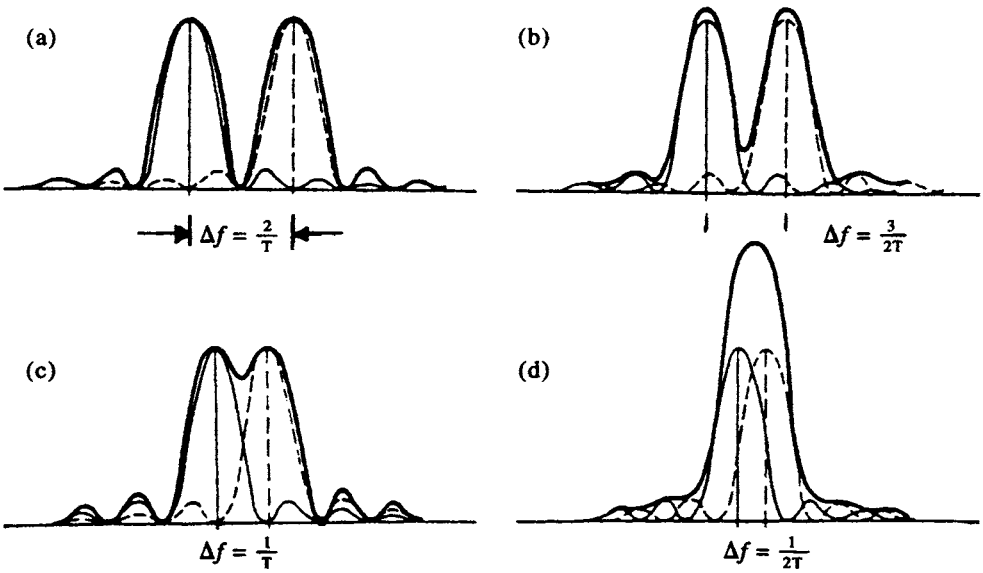


Figure 5.6.19. Resolution of spectral lines. (a, b) Well resolved; (c) just resolved; and (d) not resolved. (From Jenkins and White, 1957.)

some nonrectangular windows, the length of the data set must be increased to about $2T = 2/\Delta f$ to achieve the same frequency separation.

In a related study, Munk and Hasselman (1964) discuss the "super-resolution" of tidal frequency variability. The fact that time series of tidal heights vary at precise frequencies and have relatively large signal-to-noise ratios suggests that the traditional requirement (that a minimum record length T is required to separate tidal constituents separated by frequency difference $\Delta f = 1/T$) is "grossly incomplete". The modified resolvable frequency difference is

$$\Delta f = \frac{1}{rT}; \quad \Delta\omega = \frac{2\pi}{rT} \tag{5.6.64}$$

in which $r \equiv$ (signal level/noise level)$^{1/2}$. On this basis, the Rayleigh criterion must be considered a conservative measure of the resolution requirement for deterministic processes.

5.6.6 Smoothing spectral estimates (windowing)

The need for statistical reliability of spectral estimates brings us to the topic of spectral averaging or smoothing. As we have seen, discrete Fourier transforms provide an elegant method for decomposing a data sequence into a set of discrete spectral estimates. For a data sequence of N values, the periodogram estimate of the spectrum can have a maximum of $N/2$ Fourier components. If we use all $N/2$ components to generate the periodogram, there are only two degrees of freedom per spectral estimate, corresponding to the coefficients A_n, B_n of the sine and cosine functions for each Fourier component (see Section 5.6.3.5). Based on the assumption that data are drawn from a normally distributed random sample, we can define the confidence limits for the spectrum in terms of a chi-squared distribution, χ^2_n, where for n degrees of freedom

$$E[\chi_n^2] = \mu^2 = n, \quad E[(\chi_n^2 - \mu^2)] = \sigma^2 = 2n \tag{5.6.65}$$

Substituting $n = 2$ into these expressions, we find that the standard deviation, σ, is equal to the mean, μ, of the estimate, indicating that results based on two degrees of freedom are not statistically reliable. It is for this reason that some sort of ensemble averaging or smoothing of spectral estimates is required. The smoothing can be applied directly to the time series through convolution with a sliding averaging function or by averaging adjacent spectral estimates. A one-shot smoothing applied to the entire record increases only slightly the number of degrees of freedom per spectral estimate. In most practical applications, the full time series is broken into a series of short overlapping segments and smoothing applied to each of the overlapping segments. We then ensemble average the smoothed spectra from each segment to increase the number of degrees of freedom per spectral estimate. The more smoothing we do, the narrower the confidence limits and the greater the reliability of any observed spectral peaks. The trade-off is a loss of spectral resolution and longer processing time.

A window is a smoothing function applied to finite observations or their Fourier transforms to minimize "leakage" in the spectral domain. Convolution in the time domain and multiplication in the frequency domain are adjoint Fourier functions (see Appendix G regarding convolution). A practical window is one which allows little of the energy in the main spectral lobe to leak into the side-lobes where it can obscure and distort other spectral estimates that are present. In fact, weak signal spectral responses can be masked by higher side-lobes from stronger spectral responses. Skillful selection of tapered data windows can reduce the side-lobe leakage, although always at the expense of reduced resolution. Thus, we want a window that minimizes the side-lobes and maximizes (concentrates) the energy near the frequency of interest in the main lobe. These two performance limitations are rather troublesome when analyzing short data records. Short data occur in practice because many measured processes are of short duration or have slowly time-varying spectra that may be considered constant over only short record segments. The window is applied to data to reduce the order of the discontinuity of the boundary of the periodic extension since few harmonics will fit exactly into the length of the time series.

Signals with frequencies other than those of the basis set are not periodic in the observation window. The periodic extension of a signal, not commensurate with its natural period, exhibits discontinuities at the boundaries of the observational period. Such discontinuities are responsible for spectral contributions or leakage over the entire basis set. In the time domain, the windows are applied to the data as a multiplicative weighting (convolution) to reduce the order of the discontinuities at the boundary of the periodic extensions. The windowed data are brought to zero smoothly at the boundaries so that the periodic extensions of the data are continuous in many orders of the derivatives. The value of $Y(f)$ at a particular frequency f, say f_o, is the sum of all the spectral contributions at each f weighted by the window centered at f_o and measured at f

$$Y(f) = Y'(f) * W(f) \tag{5.6.66}$$

There exist a multitude of data windows or tapers with different shapes and characteristics ranging from the rectangular (box-car) window discussed in the previous section, to the classic Hanning and Hamming windows, to more sophisticated windows such as the Dolph–Chebyshev window. The type of window used for a given

application depends on the required degree of side-lobe suppression, the allowable widening of the central lobe, and the amount of computing one is willing to endure. We will briefly discuss several of the conventional windows plus the Kaiser–Bessel window recommended by Harris (1978).

5.6.6.1 Desired window qualities

Windows affect the attributes of a given spectral analysis method, including its ability to detect and resolve periodic waveforms, its dynamic range, confidence intervals, and ease of implementation. Spectral estimates are affected not only by the broadband noise spectrum of the data but also by narrow-band signals that fall within the bandwidth of the window. Leakage of spectral power from a narrow-band spectral component, f_o, to another frequency component, f_a, produces a bias in the amplitude and position of a spectral estimate. This bias is especially disruptive for the detection of weak signals in the presence of nearby strong signals. To reduce the bias, we need a "good" window. Although there are no universal standards for a good window, we would like it to possess the following characteristics in Fourier transform space:

(1) The central main lobe of the window (which is centered on the frequency of interest) should be as narrow as possible to improve the frequency resolution of adjacent spectral peaks in the dataset, and the first side-lobes should be greatly attenuated relative to the main lobe to avoid contamination from other frequency components. Here, the narrowness of the central lobe is measured by the positions of the −3 dB (half power points) on either side of the lobe. Retention of a narrow central lobe, while suppressing the side-lobes, is not as easy as it sounds since suppression of the side-lobes invariably leads to a broadening of the central lobe;
(2) The window should suppress the amplitudes of side-lobes at frequencies far removed from the central lobe. That is, the side-lobes should have a rapid asymptotic fall-off rate with frequency so that they leak relatively little energy into the spectral estimate at the central lobe (i.e. into the frequency of interest);
(3) The coefficients of the window should be easy to generate for multiplication in the time domain and convolution in the Fourier transform domain.

A good performance indicator (PI) for the time domain window $w(t)$ can be defined as the difference between the equivalent noise bandwidth, ENBW, and the bandwidth, BW, located between the −3 dB levels of the central lobe (Harris, 1978)

$$\mathrm{PI} = \frac{\mathrm{ENBW} - \mathrm{BW}}{\mathrm{BW}} = \frac{\dfrac{1}{\mathrm{BW}} \sum_n w^2(n\Delta t)}{\left[\sum_n w(n\Delta t)\right]^2} - 1 \tag{5.6.67}$$

where we have normalized by the bandwidth. The windows that perform well have values for this ratio (×100%) of between 4.0 and 5.5%. A summary of the figures of merit for several well-known windows is presented in Table 5.6.3. PI values are obtained using columns 4 and 5. The choice of window can be daunting; Harris lists more than 44 windows for smoothing spectral estimates.

Table 5.6.3. Windows and figures of merit. The last column gives the correlation between adjacent data segments for the specified percentage segment overlap. For completeness, we include the Tukey and Parzen windows. (From Harris, 1978)

Window	Highest side-lobe level (dB)	Side-lobe attenuation (dB/octave)	Equiv. noise BW (BINS)	3.0 dB BW (BINS)	Overlap corr. 75% 50%
Rectangle	−13	−6	1.00	0.89	0.750 0.500
Triangle	−27	−12	1.33	1.28	0.719 0.250
Hanning	−32	−18	1.50	1.44	0.659 0.167
Hamming	−43	−6	1.36	1.30	0.707 0.235
Parzen	−21	−12	1.20	1.16	0.765 0.344
Tukey $\alpha = 0.5$	−15	−18	1.22	1.15	0.727 0.364
Kaiser $\alpha = 2.0$	−46	−6	1.50	1.43	0.657 0.169
Bessel					
$\alpha = 2.5$	−57	−6	1.65	1.57	0.595 0.112
$\alpha = 3.0$	−69	−6	1.80	1.71	0.539 0.074
$\alpha = 3.5$	−82	−6	1.93	1.83	0.488 0.048

5.6.6.2 Rectangular (box-car) and triangular windows

As discussed in Section (5.6.4), a rectangular window has an amplitude of unity throughout the observation interval of duration $T = N\Delta t$, with the weighting given by

$$\begin{aligned} w(n\Delta t) &= 1, \quad n = 0, 1, \ldots, N-1 \ (\text{or } -N/2 \le n \le N/2) \\ &= 0, \quad \text{elsewhere} \end{aligned} \tag{5.6.68}$$

(Figure 5.6.20a). Using the relation $\omega T = N\theta$, where $\theta = \omega\Delta t$ and $T = N\Delta t$, the spectral window from the discrete Fourier transform (DFT) is

$$W(\theta) = T e^{-i(N-1)\theta/2} \frac{\sin(N\theta/2)}{N\theta/2} \tag{5.6.69a}$$

$$|W(\theta)|^2 = T^2 \left[\frac{\sin(N\theta/2)}{N\theta/2}\right]^2 \tag{5.6.69b}$$

(Figure 5.6.20b) where the exponential term in equation (5.6.69a) gives the phase shift of the window as a function of the frequency $\omega = \theta/\Delta t$. The function W, the Dirichlet kernal, has strong side-lobes, with the first side-lobe down only 13 dB from the main lobe. The remaining side-lobes fall off weakly at 6 dB per octave, which is the functional rate for a discontinuity (an "octave" corresponds to a factor of two in change frequency). Zeros of $W(\theta)$ occur at integer multiples of the frequency resolution, $f_1 = 1/T$, for which $N\theta/2 = \omega T/2 = \pm p\pi$. That is, where $f = \pm p/T(\pm 1/T, \pm 2/T, \ldots)$.

The triangular (Bartlett) window

$$w(n\Delta t) = \begin{cases} \dfrac{n}{(N/2)}, & n = 0, 1, \ldots, N/2 \\ \dfrac{N-n}{(N/2)}, & n = N/2, \ldots, N-1 \end{cases} \tag{5.6.70a}$$

$$= \frac{N/2 - |n|}{(N/2)}, \quad 0 \le |n| \le N/2 \tag{5.6.70b}$$

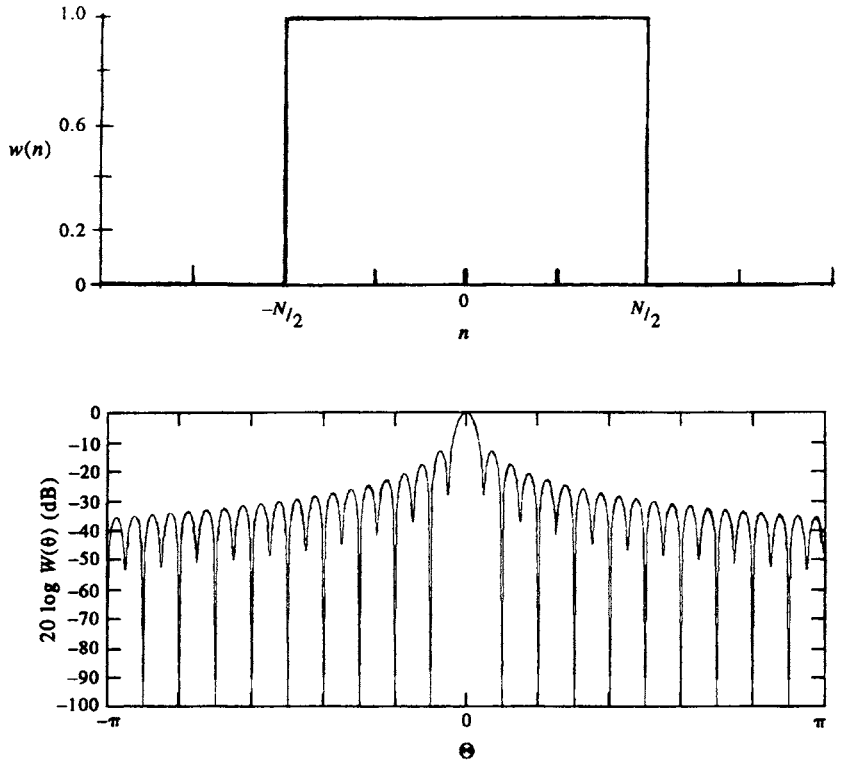


Figure 5.6.20. A box-car window for N = 41 weights. (a) Weights, w(n) = 1.0 in the time domain (−20 ≤ n ≤ 20). (b) Fourier transform of the weights, |W(θ)|, plotted as 20 log |W(θ)| where θ = ωΔt/N = 40π/N is the frequency span of the window.

(Figure 5.6.21a) has the DFT

$$W(\theta) = \frac{2T}{N} e^{-i(N-1)\theta/2} \left[\frac{\sin(N\theta/2)}{N\theta/2}\right]^2 \tag{5.6.71a}$$

$$|W(\theta)|^2 = \frac{4T^2}{N^2} \left[\frac{\sin(N\theta/2)}{N\theta/2}\right]^4 \tag{5.6.71b}$$

(Figure 5.6.21b) which we recognize as the square of the sinc function for the rectangular window. The main lobe between zero crossings is twice that of the rectangular window but the level of the first side-lobe is down by 26 dB, twice that of the rectangular window. Despite the improvement over the box-car window, the side-lobes of the triangular window are still extensive and use of this window is not recommended if other windows are available.

The Parzen window

$$w(n\Delta t) = 1.0 - |n/(N/2)|^2, \quad 0 \le |n| \le N/2 \tag{5.6.72}$$

is the squared counterpart to the Bartlett window. This is the simplest of the continuous polynomial windows and has first side-lobes down by −22 dB and falls off as $1/\omega^2$.

5.6.6.3 Hanning and Hamming windows (50% overlap)

The Hann window, or *Hanning window* as it is most commonly known, is named after the Austrian meteorologist Julius von Hann and is part of a family of trigonometric windows having the generic form $\cos^{\alpha}(n)$, where the exponent, α, is typically an integer from 1 through 4. The case $\alpha = 1$ leads to the *Tukey* (or *cosine-tapered*) *window* (Harris, 1978). As α becomes larger, the window becomes smoother, the side-lobes fall off faster and the main lobe widens. The Hanning window ($\alpha = 2$), also known as the *raised cosine* and *sine-squared* window, is defined in the time domain as

$$w(n\Delta t) = \sin^2(\pi n/N) = \tfrac{1}{2}[1.0 - \cos(2\pi n/N)], \quad n = 0, 1, \ldots, N-1 \tag{5.6.73a}$$

$$= \sin^2[\pi(n + N/2)/N]$$

$$= \tfrac{1}{2}[1.0 - \cos[2\pi(n + N/2)/N]], \quad n = -N/2, \ldots, N/2 \tag{5.6.73b}$$

(Figure 5.6.22a) which is a continuous function with a continuous first derivative. The DFT of this weighting function is

$$W(\theta) = \tfrac{1}{2}D(\theta) + \tfrac{1}{4}[D(\theta - \theta_1) + D(\theta + \theta_1)] \tag{5.6.74}$$

(Figure 5.6.22b) where $\theta_1 = 2\pi/N$ and

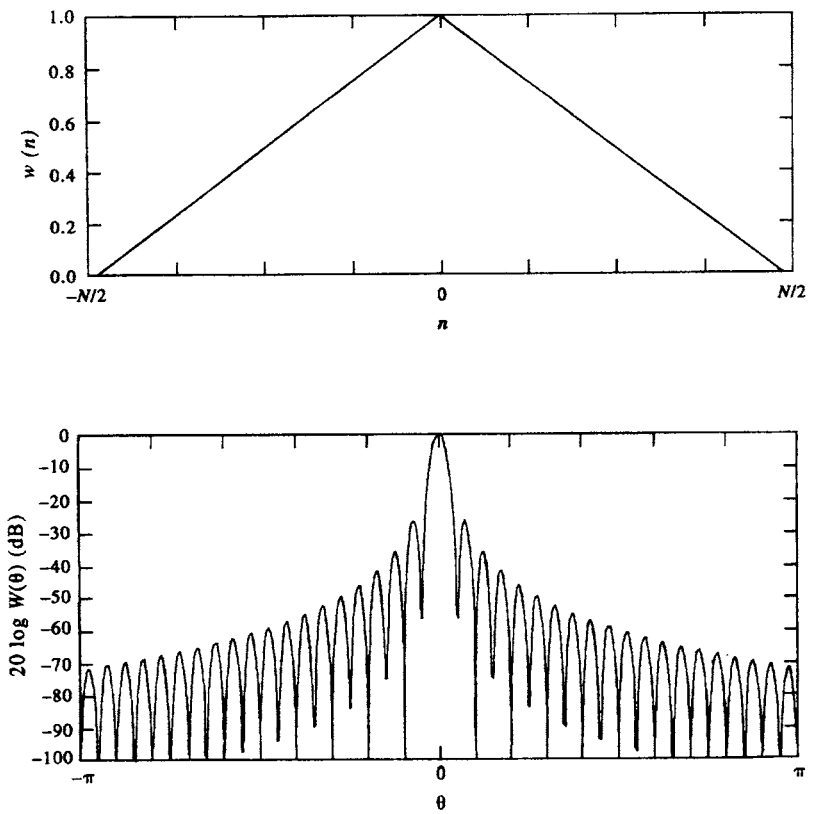


Figure 5.6.21. The triangular (Bartlett) window for N = 51 weights. (a) Weights, w(n) in the time domain ($-20 \leq n \leq 20$). (b) Fourier transform of the weights, $|W(\theta)|$, plotted as 20 log $|W(\theta)|$ (cf. Figure 5.6.20).

$$D(\theta) = Te^{i\theta/2} \frac{\sin(N\theta/2)}{N\theta/2} \tag{5.6.75}$$

is the standard function (Dirichlet kernal) obtained for the rectangular and triangular windows. Thus, the window consists of the summation of three sinc functions (Figure 5.6.22c), one centered at the origin, $\theta = 0$, and two other translated Dirichlet kernals having half the amplitude of the main kernal and offset by $\theta = \pm 2\pi/N$ from the central lobe. There are several important features of the window response $W(\theta)$. First of all, the functions D are discrete and defined only at points which are multiples of $2\pi/N$, which also correspond to the zero crossings of the central function, $D(\theta)$. Secondly, for all of zero crossings except those at $\theta_{\pm 1} = \pm 2\pi/N$, the translated functions also have zero crossings at multiples of $2\pi/N$. As a result, only values at $-2\pi/N$, 0, and $+2\pi/N$ contribute to the window response. It is the widening of the main lobes of the translated functions that causes them to be nonzero at the first zero-crossings of the central function. Lastly, because the translated functions are out of phase with the central function, they tend to cancel the side-lobe structure. The first side-lobe is down by 32 dB from the main lobe. The remaining side-lobes diminish as $1/\omega^3$ or at about -18 dB per octave.

An attractive aspect of the Hanning window is that smoothing in the frequency domain can be accomplished using only three convolution terms corresponding to θ_o, $\theta_{\pm 1}$. The Hanning-windowed Fourier transform Y_H for the spectral frequency, f_k, is then obtained from the raw spectra Y for the frequencies f_k and the two adjoining frequencies f_{k-1}, f_{k+1}; that is

$$Y_H(f_k) = \tfrac{1}{2}\{Y(f_k) - \tfrac{1}{2}[Y(f_{k-1}) + Y(f_{k+1})]\} \tag{5.6.76}$$

The transform $Y(f_k)$ has been rectangular-windowed by the act of collecting the data but is "raw" in the sense that no additional smoothing has been applied. Other processing advantages of the Hanning window are discussed by Harris (1978). Since the squares of the weighting terms $(1/2)^2 + (1/4)^2 + (1/4)^2 = 3/8$, the total energy will be reduced following the application of the Hanning window. To compensate, the amplitudes of the Fourier transforms $Y_H(f)$ should be multiplied by $\sqrt{(8/3)}$ prior to computation of the spectra. Specifically

$$Y_H(f_k) = \Delta t(8/3)^{1/2} \sum_{n=0}^{N-1} y_n[1 - \cos(2\pi n/N)]e^{-i2\pi kn/N} \tag{5.6.77}$$

where $f_k = k/(N\Delta t)$.

The *Hamming window* is a variation on the Hanning window designed to cancel the first side-lobes. To accomplish this, the relative sizes of the three Dirichlet kernels are adjusted through a parameter, γ where

$$w(n\Delta t) = \gamma + (1 - \gamma)\cos(2\pi n/N)], \quad n = -N/2, ..., N/2 \tag{5.6.78a}$$

$$W(\theta) = \gamma D(\theta) + \tfrac{1}{2}(1 - \gamma)[D(\theta - 2\pi/N) + D(\theta + 2\pi/N)] \tag{5.6.78b}$$

Perfect cancellation of the first side-lobes (located at $\theta_1 = 2.5\pi/N$) occurs when $\gamma = 25/46 \approx 0.543478$. Taking $\gamma = 0.54$ leads to near-perfect cancellation at

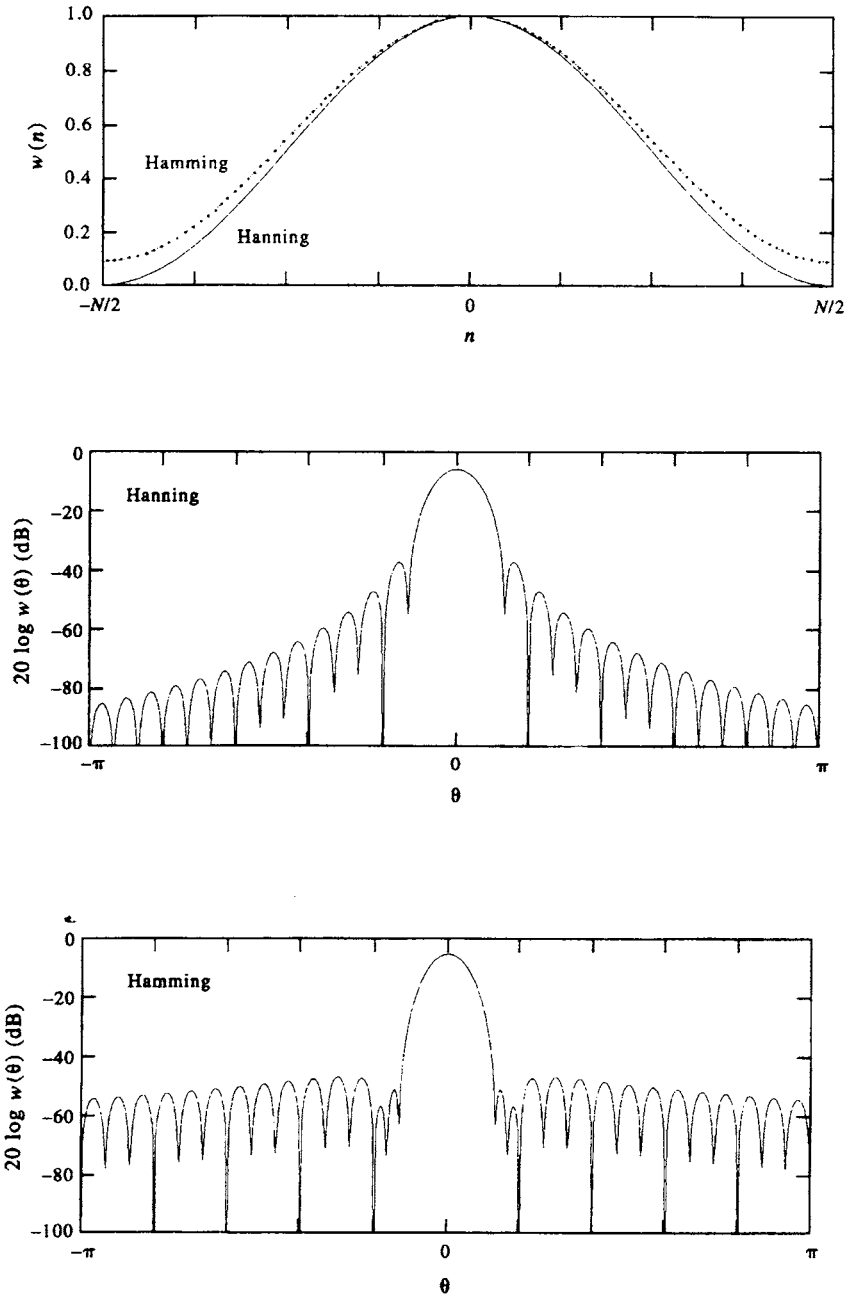


Figure 5.6.22. The Hanning and Hamming windows for $N = 41$ weights. (a) Weights, $w(n)$, $(-20 \leq n \leq 20)$. (b) Fourier transform of the weights, $|W(\theta)|$, plotted as 20 log $|W(\theta)|$ (cf. Figure 5.6.20). The response functions have not been re-scaled.

$\theta_1 = 2.6\pi/N$ and a marked improvement in side-lobe level. The Hamming window is defined as

$$w(n\Delta t) = 0.54 + 0.46\cos(2\pi n/N), \quad n = -N/2, ..., N/2 \tag{5.6.79}$$

and has a spectral distribution similar to that of the Hanning window with more "efficient" side-lobe attenuation. The highest side-lobe levels of the Hanning window occur at the first side-lobes and are down by 32 dB from the main lobe. For the Hamming window, the first side-lobe is highly attenuated and the highest side-lobe level (the third side-lobe) is down by 43 dB. To compensate for the filter, the amplitudes of the Fourier transforms $Y_{\text{Ham}}(f)$ should be multiplied by $\sqrt{(5/2)}$ prior to computation of the spectra. On a similar note, if you are going to use any of the windows in this section to calculate running mean time series, make sure each estimated value is divided by the sum of the weights used, $\sum_N w_n$.

5.6.6.4 Kaiser–Bessel window (75% overlap)

Harris (1978) identifies the Kaiser–Bessel window as the "top performer" among the many different types of windows he considered. Among other factors, the coefficients of the window are easy to generate and it has a high equivalent noise bandwidth, one of the criteria used to separate good and bad windows. The trade-off is increased main-lobe width for reduced side-lobe levels. In the time domain the filter is defined in terms of the zeroth-order modified Bessel functions of the first kind.

$$w(n\Delta t) = \frac{I_o(\pi\alpha\Omega)}{I_o(\pi\alpha)}, \quad 0 \le |n| \le N/2 \tag{5.6.80}$$

where the argument $\Omega = [1.0 - (2n/N)^2]^{1/2}$ and

$$I_o(x) = \sum_{k=0}^{\infty}\left[\frac{(x/2)^k}{k!}\right]^2 \tag{5.6.81}$$

The parameter $\pi\alpha$ is half of the time-bandwidth product, with α typically having values 2.0, 2.5, 3.0, and 3.5. The transform is approximated by

$$W(\theta) \approx [N/I_o(\pi\alpha)]\frac{\sinh\{[\pi^2\alpha^2 - (N\theta/2)^2]^{1/2}\}}{\{[\pi^2\alpha^2 - (N\theta/2)^2]^{1/2}\}} \tag{5.6.82}$$

Plots of the weighting function w and the DFT for W are presented in Figure 5.6.23 for two values of the parameter $\alpha(= 2.0, 3.0)$. The modified Bessel function I_o is defined as follows.

For $|x| \le 3.75$

$$I_o(x) = \{[\{[(4.5813 \times 10^{-3}Z + 3.60768 \times 10^{-2})Z + 2.659732 \times 10^{-1}]Z + 1.2067492\}Z + 3.0899424]Z + 3.5156229\}Z + 1.0 \tag{5.6.83a}$$

where for real x

$$Z = (x/3.75)^2 \tag{5.6.83b}$$

For $|x| > 3.75$

$$I_o(x) = \exp(|x|)/|x|^{1/2}\{[(\{[\{[(3.92377 \times 10^{-3}Z - 1.647633 \times 10^{-2})Z + 2.635537 \times 10^{-2}]Z - 2.057706 \times 10^{-2}\}Z + 9.16281 \times 10^{-3}]Z - 1.57565 \times 10^{-3}\}Z + 2.25319 \times 10^{-3})Z + 1.328592 \times 10^{-2}]Z + 3.9894228 \times 10^{-1}\} \tag{5.6.83c}$$

where

$$Z = 3.75/|x| \tag{5.6.83d}$$

The usefulness of the Kaiser–Bessel window is nicely illustrated by Figure 5.6.24. Here, we compare the average spectra (in cm^2/cpd) obtained from a year-long record of hourly coastal sea level following application of a rectangular window (the worst possible window) and a Kaiser–Bessel window (the best possible window) to a series of overlapping data segments. In each case, the window length is 42.7 days and there are $K = 32$ degrees of freedom per spectral estimate, corresponding to roughly 16 separate spectral estimates for 50% window overlaps. Both windows preserve the strong spectra peaks within the tidal frequency bands centered at 1, 2, and 3 cpd. However, unlike the rectangular window, the Kaiser–Bessel window results in little energy leakage from the tidal bands to adjacent frequency bands. The high spectral levels at periods shorter than about two days ($f > 0.5$ cpd) in the nontidal portion of the rectangularly windowed spectra is an artifact of the window. The slightly better ability of the rectangular window to resolve frequency components within the various tidal bands is outweighed by the high contamination of the spectrum at nontidal frequencies.

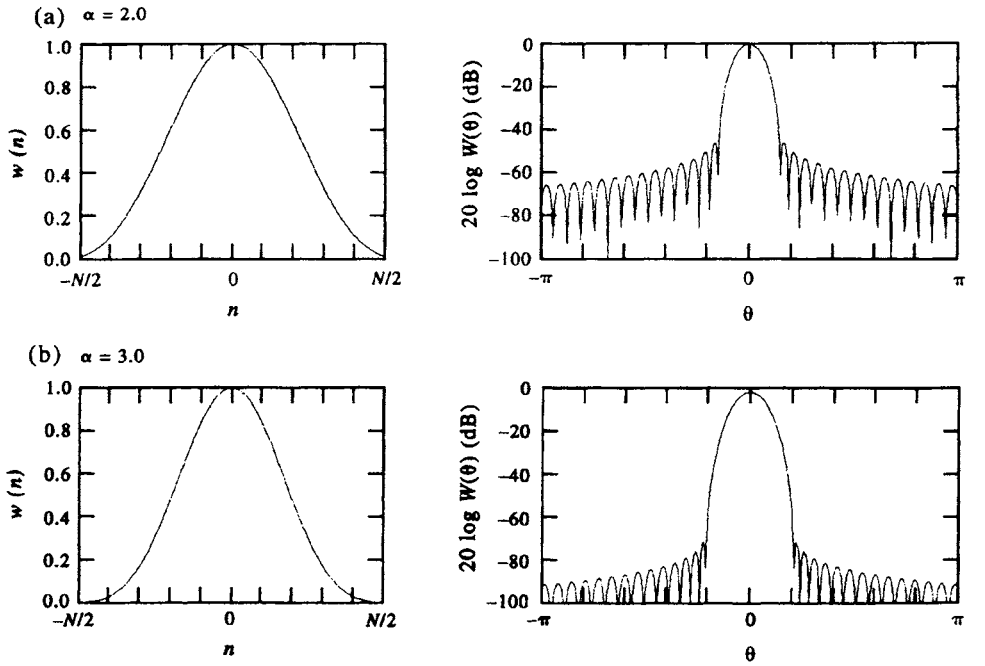


Figure 5.6.23. The Kaiser–Bessel window for $N = 51$ weights and $\alpha = 2.0$ and 3.0. (a) Weights, $w(n)$, $(-20 \leq n \leq 20)$. (b) Fourier transform of the weights, $|W(\theta)|$, plotted as 20 log $|W(\theta)|$ (cf. Figure 5.6.20). (From Harris, 1978.)

5.6.7 Smoothing spectra in the frequency domain

As we noted earlier, each spectral estimator for a random process is a chi-squared function with only two degrees of freedom (DOF). Because of this minimal number of degrees of freedom, some sort of smoothing or filtering is needed to increase the statistical significance of a given spectral estimate. The windowing approach described in the previous section, in which we partitioned the time series into a series of shorter overlapping segments, is one of a number of computational methods used to smooth (average) spectral estimates.

5.6.7.1 Band averaging

For a time series consisting of N data points, one of the simplest forms of smoothing is to use the discrete Fourier transform or fast Fourier transform to calculate individual spectral estimates for the maximum number of frequency bands ($N/2$) and then average together adjacent spectral estimates. The resultant spectral estimate is assigned to the mid-point of the average. Thus, we could average bands 1, 2, and 3, to form a single spectral estimate centered at band 2, then bands 4, 5, and 6 to form an estimate centered at band 5, and so on. It is often useful in this type of *frequency band averaging* to use an odd-numbered smoother so that the center point is easily defined. In particular, if we were to average groups of three adjacent (and different) bands to form each estimate, the number of degrees of freedom per estimate would increase

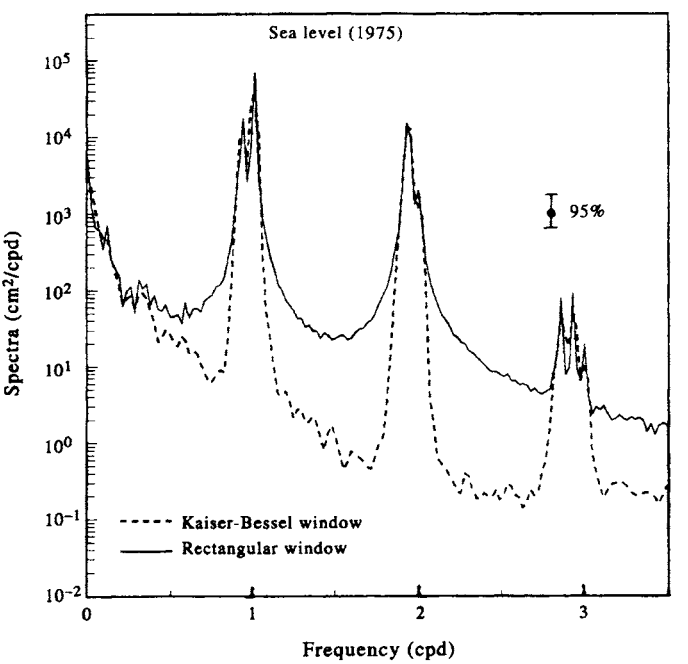


Figure 5.6.24. Spectra (cm^2/cpd) of the hourly coastal sea-level height recorded at Victoria, British Columbia during 1975 following windowing (number of hourly samples, $N = 8750$). Linear frequency. Solid line: Rectangular window. Dashed line: Kaiser–Bessel window with $\alpha = 3$. Both windows have a length of 1024 h (= 42.67 days) and there are DOF = 32, using a total of 16 50% overlapping data segments. The tidal peak centered at 3 cpd results from nonlinear interactions within the semidiurnal frequency band. Vertical line is the 95% level of confidence. (Courtesy, A. Rabinovich.)

from 2 to 6. In the case of the Blackman–Tukey procedure, an alternative method is to use bigger lag steps in the computation of the autocovariance function before its transform is taken. This is functionally equivalent to smoothing by averaging together the individual spectral estimates.

5.6.7.2 Block averaging

As noted earlier, a common smoothing technique is to segment the time series (of length N) into a series of short, equal-length segments of length N_s (where $N = KN_s$, and K is a positive integer). Spectra are then computed for each of the K segments and the spectral values for each frequency band then *block averaged* to form the final spectral estimates for each frequency band. If there is no overlap between segments, the resulting degrees of freedom for the composite spectrum will be $2K$. This assumes that the individual sample spectra have not been windowed and that each spectral estimate is a chi-squared variable with 2 degrees of freedom. Since the frequency resolution of a time series is inversely proportional to its length, the major difficulty with this approach is that the shorter time series have fewer spectral values than the original record over the same Nyquist frequency range. In other words, the maximum resolvable frequency $1/2\Delta t$ remains the same since Δt is unchanged, but the frequency spacing between adjacent spectral estimates is increased for the short segments because of the reduced record lengths.

However, by not overlapping adjacent segments, we could be overly conservative in our estimate of the number of degrees of freedom. For that reason, most analysts overlap adjacent segments by 30–50% so that more uniform weighting is given to individual data points. The need for overlapping segments is necessary when a window is applied to each individual segment prior to calculation of the spectra. The effect of the window is to reduce the effective length of each segment in the time domain so that, for some sharply defined windows such as the Kaiser–Bessel window, even adjoining segments with 50% overlap can be considered independent time series for spectral analysis. As in Figure 5.6.24, The degrees of freedom of the periodograms averaged together is $4K$, rather than $2K$ for the nonoverlapping segments. Consideration must be given to the correlation among individual estimates (the greater the overlap the higher the correlation). Nuttall and Carter (1980) report that 92% of the maximum number of equivalent degrees of freedom can be achieved for a Hanning window which uses 50% overlap. Clearly, we must sacrifice something to gain improved statistical reliability. That "something" is a loss of frequency resolution due to the broad central lobe that accompanies windows with negligible side-lobes.

As an example, consider the spectrum of a 1-min sampled time series $y(t) = A\cos(2\pi ft) + \varepsilon(t)$ of length 512 min composed of Gaussian white noise $\varepsilon(t)(|\varepsilon| \leq 1)$ and a single cosine component of amplitude, A, and frequency f = 0.23 cpmin (period $T = 1/f = 4.3$ min). The magnitude of the deterministic component, A, is five times the standard deviation of the white noise signal and $V[\varepsilon] = (1/\sqrt{2})$ cm^2. The raw periodogram (Figure 5.6.25a) reveals a large narrow peak at the frequency (0.23 cpmin) of the single cosine term plus a large number of smaller peaks associated with the white noise oscillations. In this case, there has been no spectral smoothing and the resultant spectral estimates are chi-squared functions with 2 degrees of freedom. The variances of the spectral peaks are as large as the peaks themselves. If we average together three adjacent spectral components (Figure 5.6.25b), we obtain a much smoother spectrum, $S(f)$. Here, $S_i = S(f_i)$ is defined by $S_i =$

$1/3[S(f_{i-1}) + S(f_i) + S(f_{i+1})]$, $S_{i+3} = 1/3[S(f_{i+2}) + S(f_{i+3}) + S(f_{i+4})]$, and so on. Each of the new spectral estimates now has six degrees of freedom instead of only two. The bottom two panels in this figure show what happens if we increase the number of frequency bands averaged together to seven (Figure 5.6.25c) and then to 15 (Figure 5.6.25d). Note that, with increasing degrees of freedom (DOF), our confidence in the existence of a spectral peak increases but delineation of the peak frequency decreases. With increasing DOF, there is increased smoothing of all spectral peaks (see also Figure 5.6.24). The same effect can be achieved by operating on the autocovariance function rather than on the Fourier spectral estimates. In particular, a spectrum similar to Figure 5.6.25(a) is obtained using the autocovariance transform method on the time series $y(t)$ for a time lag of 1 min (the sampling interval). If we apply a lag of 3 min in computing the autocovariance transform, we obtain a spectrum similar to Figure 5.6.25(b), and so on. Any differences between the two methods will be due to computational uncertainties.

To determine the number of degrees of freedom for any block averaging, we define the normalized standard error $\varepsilon(\tilde{G})$ of the one-sided spectrum, $\tilde{G}_{yy}(f)$, of the time series $y(t)$ of length $T = N\Delta t$, as

$$\varepsilon[\tilde{G}_{yy}(f)] = \frac{V[\tilde{G}_{yy}(f)]^{1/2}}{G_{yy}(f)} \tag{5.6.84}$$

where $V[\tilde{G}]$ is the variance of $\tilde{G}$, the tilde ($\sim$) denotes the raw estimate of the time series, and

$$\tilde{G}_{yy}(f)/G_{yy}(f) = \chi_2^2/2 \tag{5.6.85}$$

is a chi-square variable with $n = 2$ degrees of freedom. For the narrowest possible resolution $\Delta f = 1/T$, we have

$$\varepsilon[\tilde{G}_{yy}(f)] = \frac{(2n)^{1/2}}{n} = (2/n)^{1/2} \tag{5.6.86}$$

For maximum resolution, $n = 2$ and so $\varepsilon(\tilde{G}) = 1$, giving the not-so-useful result that the standard deviation of the estimate is as large as the estimate itself. If, on the other hand, we average the spectral estimates for each frequency for the maximum resolution spectra using a total of N_s separate and independent record segments of length T_s (where $T = N_s \cdot T_s$) we find

$$\tilde{G}_{yy}(f) = \frac{2}{N_s T_s} \sum_{i=1}^{N_s} |Y_i(f_i, T_s)|^2 \tag{5.6.87}$$

so that

$$\varepsilon[\tilde{G}_{yy}(f)] = (2n/2N_s)^{1/2} = (1/N_s)^{1/2} \tag{5.6.88}$$

The resolution (effective) bandwidth is $b_e = N_s/T = 1/T_s$. Since the first estimate gives two degrees of freedom per spectral band, this gives $2N_s$ degrees of freedom per frequency band.

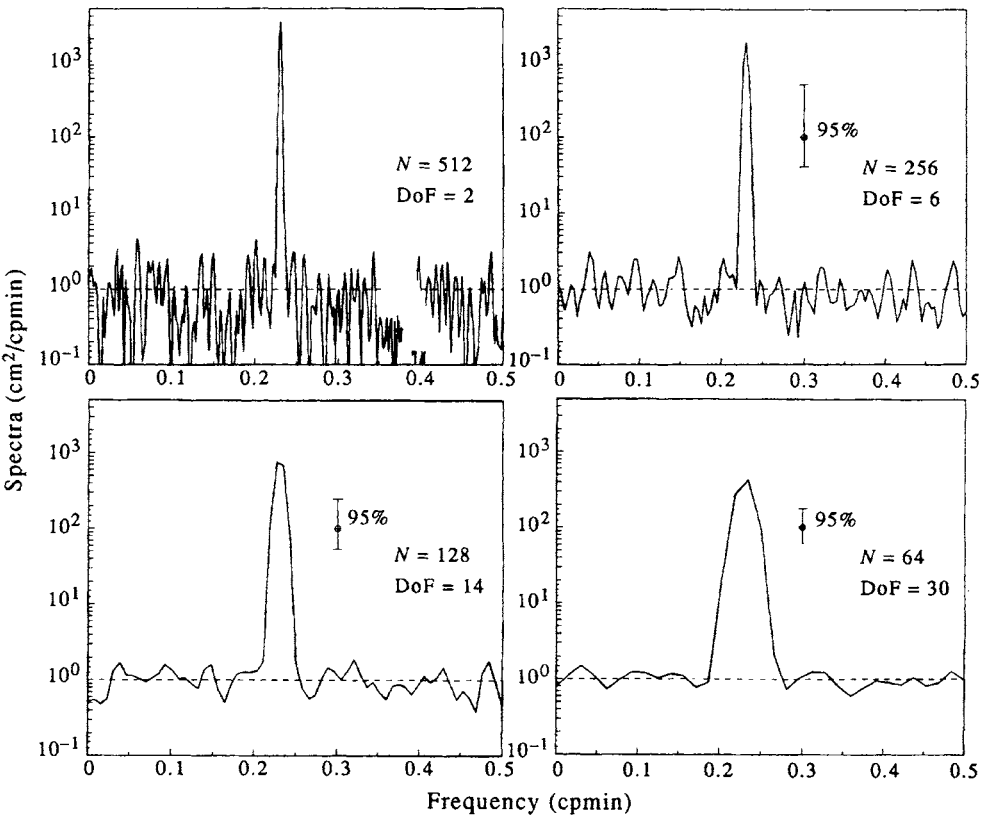


Figure 5.6.25. Periodogram power spectral estimates for a time series composed of Gaussian white noise and a single cosine constituent with a frequency of 0.23 cpmin and amplitude five times that of the white noise component. N = number of spectral bands and vertical lines are the 95% confidence intervals. (a) Raw (unsmoothed) periodogram, with DOF = 2; (b) smoothed periodogram, by averaging three adjacent spectral estimates such that DOF = 6; (c) as with (b) but for seven frequency bands, and DOF = 14; as with (b) but for 15 frequency bands, DOF = 30.

5.6.8 Confidence intervals on spectra

We can generalize equation (5.6.85) by noting that the ratio of the estimated spectrum and the expected values of the spectrum

$$\frac{\nu \tilde{G}_{yy}(f)}{G_{yy}(f)} = \chi^2_\nu \tag{5.6.89}$$

is distributed as a chi-square variable with ν degrees of freedom. It then follows that

$$P\left[\chi^2_{\alpha/2,\nu} < \frac{\nu \tilde{G}_{yy}(f)}{G_{yy}(f)} < \chi^2_{1-\alpha/2,\nu}\right] = 1 - \alpha \tag{5.6.90}$$

where

$$P[\chi^2_\nu \leq \chi^2_{\alpha/2,\nu}] = \alpha/2 \tag{5.6.91}$$

Thus, the true spectrum, $G_{yy}(f)$, is expected to fall into the interval

$$\frac{\nu\tilde{G}_{yy}(f)}{\chi^2_{1-\alpha/2,\nu}} < G_{yy}(f) < \frac{\nu\tilde{G}_{yy}(f)}{\chi^2_{\alpha/2,\nu}} \tag{5.6.92}$$

with $(1-\alpha)100\%$ confidence. In this form, the confidence limit applies only to the frequency f and not to other spectral estimates. We further point out that the degrees of freedom, ν, in the above expressions are different for windowed and nonwindowed time series. For windowed time series, we need to use the "equivalent" degrees of freedom, as presented in Table 5.6.4 for some of the more commonly used windows.

Another way to view these arguments is to equate $\tilde{G}_{yy}(f)$ with the measured standard deviation, $s^2(f)$, of the spectrum and $G_{yy}(f)$ with the true variance, $\sigma^2(f)$. Then

$$\frac{(\nu-1)s^2(f)}{\chi^2_{1-\alpha/2,\nu}} < \sigma^2(f) < \frac{(\nu-1)s^2(f)}{\chi^2_{\alpha/2,\nu}} \tag{5.6.93}$$

If spectral peaks fall outside the range (5.6.92) then to the $(1-\alpha)100\%$ confidence level they cannot have occurred by chance. The confidence levels are found by looking up the values for $\chi^2_{1-\alpha/2,\nu}$ and $\chi^2_{\alpha/2,\nu}$ in a chi-square table, then calculating the intervals based on the observed standard deviation, s. (Confidence limits on spectral coherency functions are given in Section 5.8.6.1.)

5.6.8.1 Confidence intervals on a logarithmic scale

The confidence intervals derived above apply only to individual frequencies, f. This results from the fact that the confidence interval is determined by the value $G_{yy}(f)$ of the spectral estimate and will be different for each spectral estimate. It would be convenient if we could have a single confidence interval that applies to all of the spectral values at all frequencies. To obtain such a confidence interval, we transform the spectrum using the $\log_{10}$ function. Transforming the above confidence limits we have

$$\log[\tilde{G}_{yy}(f)] + \log[\nu/\chi^2_{1-\alpha/2,\nu}] \leq \log[G_{yy}(f)] \leq \log[\tilde{G}_{yy}(f)] + \log[\nu/\chi^2_{\alpha/2,\nu}] \tag{5.6.94}$$

or

$$\log[\nu/\chi^2_{1-\alpha/2,\nu}] \leq \log[G_{yy}(f)] - \log[\tilde{G}_{yy}(f)] \leq \log[\nu/\chi^2_{\alpha/2,\nu}] \tag{5.6.95}$$

When the estimated spectrum is plotted on a log scale, a single vertical confidence

Table 5.6.4. Equivalent degrees of freedom for spectra calculated using different windows. N is the number of data points in the time series and M is the half-width of the window in the time domain. (From Priestley, 1981). $N \neq M$ for the truncated periodogram

Type of window	Equivalent degrees of freedom
Truncated periodogram	N/M
Bartlett window	$3N/M$
Daniell window	$2N/M$
Parzen window	$3.708614(N/M)$
Hanning window	$(8/3)(N/M)$
Hamming window	$2.5164\ (N/M)$

interval is determined for all frequencies by the upper and lower bounds in the above expression (Figure 5.6.26a). The spectral estimate $G_{yy}(f)$ itself is no longer a part of the confidence interval. This aspect, together with the fact that most spectral amplitudes span many orders of magnitude, is a principal reason for presenting spectra as log values. If larger numbers of spectral estimates are averaged together at higher frequencies (i.e. ν is increased), the confidence interval narrows with increasing frequency (Figure 5.6.26b). Note that the length of the confidence interval is longer above the central point than below.

5.6.8.2 Fidelity and stability

The general objective of all spectral analysis is to estimate the function $G_{yy}(f)$ as accurately as possible. This involves two basic requirements:

(1) That the mean smoothed spectrum, $\tilde{G}_{yy}(f)$, be as close as possible to the actual spectrum $G_{yy}(f)$. That is, the bias

$$B(f) = G_{yy}(f) - \tilde{G}_{yy}(f) \tag{5.6.96}$$

should be small. If this is true for all frequencies, then $\tilde{G}_{yy}(f)$ is said to reproduce $G_{yy}(f)$ with high *fidelity*.

(2) For a time series of length T that has been segmented into M pieces for spectral estimation, the variance of the smoothed spectral estimator for bandwidth b_1 is

$$V[\tilde{G}_{yy}(f)] \approx \frac{(M/b_1)}{T}[G_{yy}(f)]^2 \tag{5.6.97}$$

and should be small. If this is true, the spectral estimator is said to have high *stability*.

5.6.9 Zero-padding and prewhitening

For logistical reasons, many of the time series that oceanographers collect are too short for accurate definition of certain spectral peaks. The frequency resolution $\Delta f = 1/T$ for a record of length T may not be sufficient to resolve closely spaced spectral components. Also, discrete points in the computed spectrum may be too widely spaced to adequately delineate the actual frequency of the spectral peaks. Unfortunately, the first problem—that of trying to distinguish waveforms with nearly the same frequency—can only be solved by collecting a longer time series; i.e. by increasing T to sharpen up the frequency resolution f of the periodogram. However, the second problem—that of locating the frequency of a spectral peak more precisely—can be addressed by padding (extending) the time series with zeros prior to Fourier transforming. Transforming the data with zeros serves to refine the frequency scale through interpolation between power spectral density estimates within the Nyquist interval $-f_N \leq f \leq f_N$. That is, additional frequency components are added between those that would be obtained with a nonzero-padded transform. Adding zeros helps fill in the shape of the spectrum but in no case is there an improvement in the fundamental frequency resolution. *Zero-padding* is useful for: (1) smoothing the appearance of the periodogram estimates via interpolation; (2) resolving potential ambiguities where the frequency difference between line spectra is greater than the fundamental frequency resolution; (3) helping define the exact frequency of spectral peaks by reducing the "quantization" accuracy error; and (4)

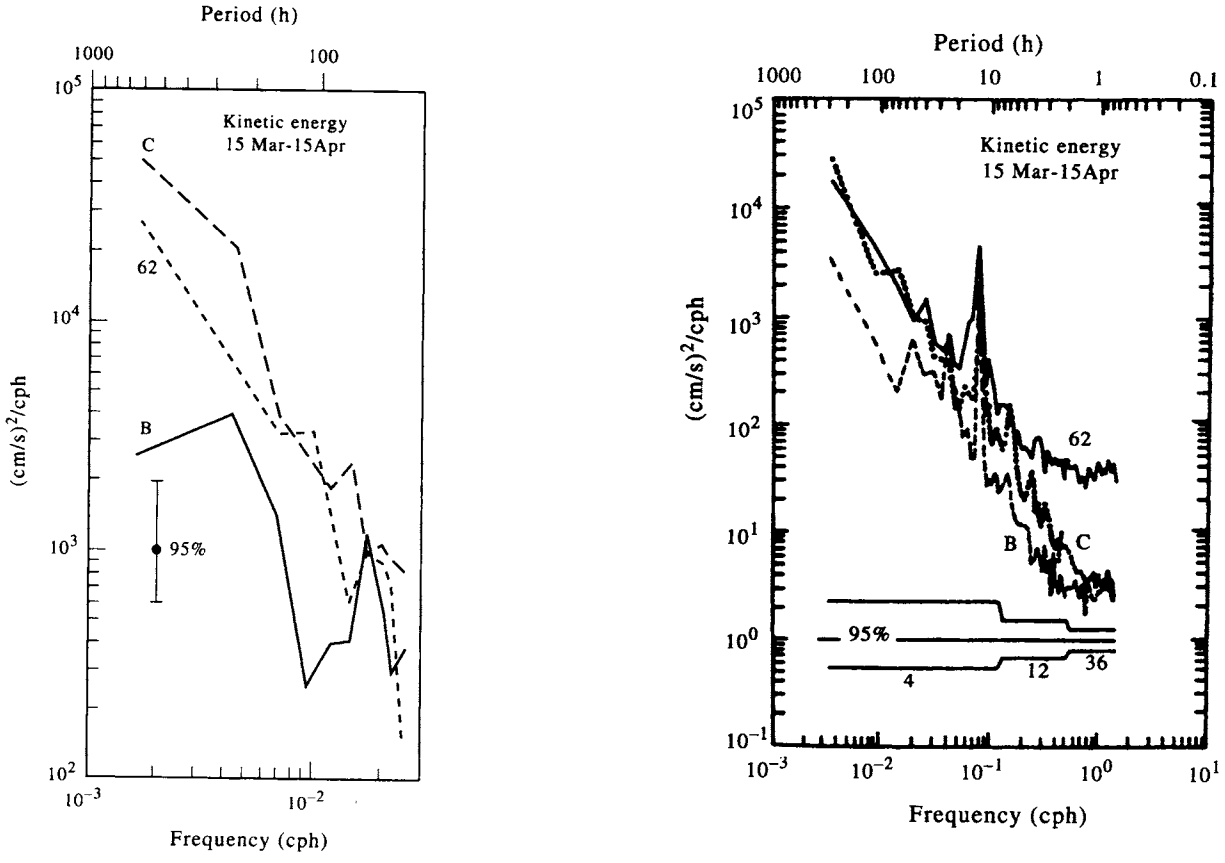


Figure 5.6.26. Confidence intervals for current velocity spectra at 50-m depth for three locations (B,C, and 62) on the northeast Gulf of Alaska shelf (59.5°N, 142.2°W), 15 March to 15 April 1976. (a) 95% interval for the low-pass filtered currents. The single vertical bar applies to all frequencies; (b) 95% interval for unfiltered records. Confidence interval narrows at higher frequencies with the increased number of degrees of freedom (from 4 to 36) used in selected frequency ranges. (Adapted from Muench and Schumacher, 1979.)

extend the number of samples to an integer power of 2 for FFT analysis. An example of how zero-padding improves the spectral resolution of a simple digitized data set is provided in Figure 5.6.27. We again emphasize that increased zero-padding helps locate the frequency of discernible spectral peaks, in this case the peaks of the sin x/x function, but cannot help distinguish closely spaced frequency components that were unresolved by the original time series prior to padding.

Prewhitening is a filtering or smoothing technique used to improve the statistical reliability of spectral estimates by reducing the leakage from the most intense spectral components and low-frequency components of the time series that are poorly resolved. To reduce the biasing of these components, the data are smoothed by a window whose spectrum is inversely proportional to the unknown spectrum being considered. Within certain frequency bands, the spectrum becomes more uniformly distributed and approaches that of white noise. Information on the form of the window necessary to construct the white spectrum must be available prior to the application of the smoothing. In effect, the time series $y(n\Delta t)$ is filtered with the weighting function $h(n\Delta t)$ such that the output is

$$y'(n\Delta t) = h(n\Delta t) \cdot (n\Delta t) \tag{5.6.98}$$

has a nearly white spectrum. Once the spectrum $S'_y(\omega)$ is determined, the desired spectrum is derived directly as

$$S_y(\omega) = \frac{S'_y(\omega)}{|H(\omega)|^2} \tag{5.6.99}$$

The best aspects of the parametric and nonparametric spectral techniques can be combined if a parametric model is used to prewhiten the time series prior to the application of a smoothed periodogram analysis. In most prewhitening situations, one is limited to using the first-difference filter in which the current data value is subtracted from the next value multiplied by some weighting coefficient, $0 \leq \alpha \leq 1$. That is $y'(t) = y(t) - \alpha y(t + \Delta t)$. The weighting coefficient can be taken as equal to the correlation coefficient of the initial data series with a shift of one time step, Δt. The filter suppresses low frequencies and stresses high frequencies and has a transform

$$H(f) = [1 - \alpha\, e^{-i2\pi f \Delta t}]^2 = 1 - 2\alpha \cos(2\pi f \Delta t) + \alpha^2 \tag{5.6.100}$$

Prewhitening reduces leakage and increases the effectiveness of frequency averaging of the spectral estimate (reduces the random error). The reduced leakage gives rise to a greater dynamic range of the analysis and allows us to examine weak spectral components. Notice that, if $Y(f)$ is the Fourier transform of $y(t)$, then the Fourier transform of $y'(t)$ is

$$Y'(\omega) = \int_t y'(t) e^{-i\omega t} dt \approx \omega \cdot Y(\omega) \tag{5.6.101}$$

so that *first differencing* is like a linear high-pass filter with amplitude $|H(\omega)| = |\omega|$. This effect shows up quite well in the processing of satellite-tracked drifter data. Spectra of the drifter positions (longitude, $x(t)$; latitude, $y(t)$) as functions of time, t, are generally "red" whereas the spectra of the corresponding drifter velocities (zonal, $u = \Delta x/\Delta t$; meridional, $v = \Delta y/\Delta t$) are considerably "whiter" (Figure 5.6.28).

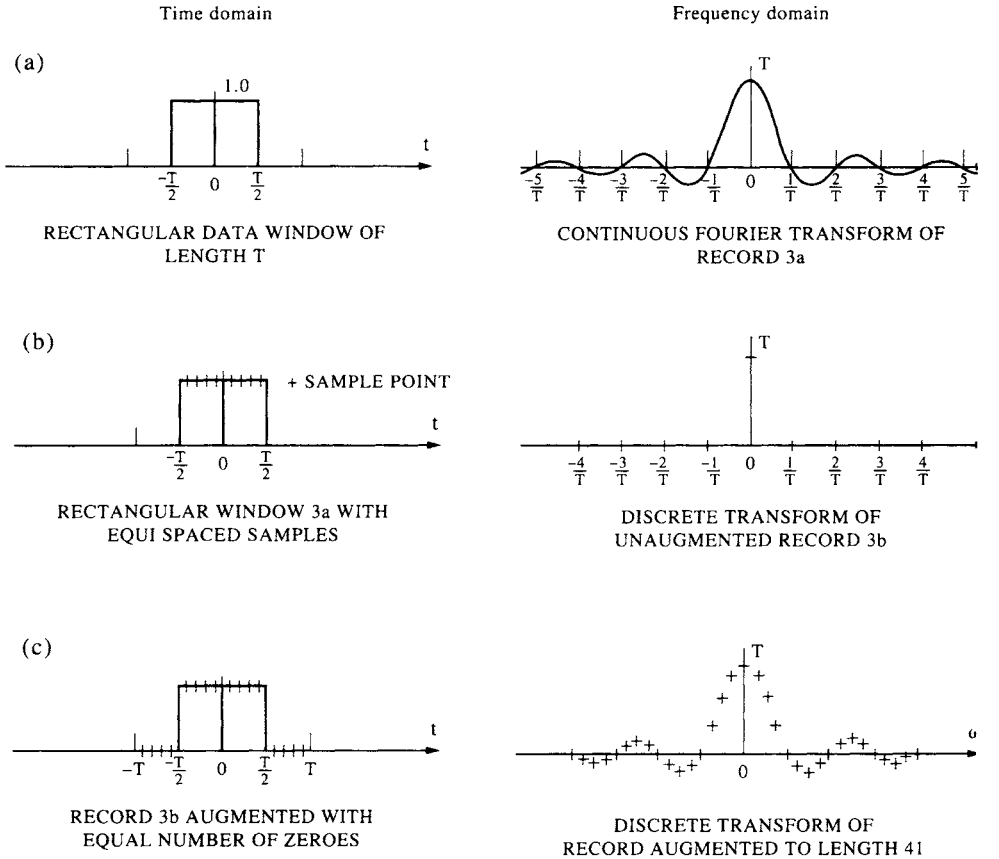


Figure 5.6.27. Use of zero padding to improve the delineation of spectral peaks. (a) A continuous box-car window of length T and its continuous Fourier transform; (b) a discrete sample of (a) at equally spaced sampling intervals and its discrete Fourier transform; (c) same as (b) but with zero padding of 2T data points. (From Henry and Graefe, 1971.)

5.6.10 Spectral analysis of unevenly spaced time series

Most discrete oceanographic time-series data are recorded at equally spaced time increments. However, some situations arise where the recorded data are spaced unevenly in time or space. For example, positional data obtained from satellite-tracked drifters are sampled at irregular time intervals due to the eastward progression in the swaths of polar-orbiting satellites and to the advection of the drifters by surface currents. Repeated time-series oceanic transects are typically spaced at irregular intervals due to the vagaries of ship scheduling and weather. In addition, instrument failure and data drop-outs generally lead to "gappy", irregularly spaced time series.

As noted in Section 3.17, a common technique for dealing with irregularly sampled or gappy data is to interpolate data values to a regular grid. This works well as long as there are not too many gaps and the gaps are of short duration relative to the signals of interest. Long data gaps can lead to the creation of erroneous low-frequency oscillations in the data at periods comparable to the gap lengths. Only for the least-squares method for harmonic analysis described in Section 5.5 is unevenly sampled data perfectly acceptable. Vaníček (1971), Lomb (1976) and others have devised a

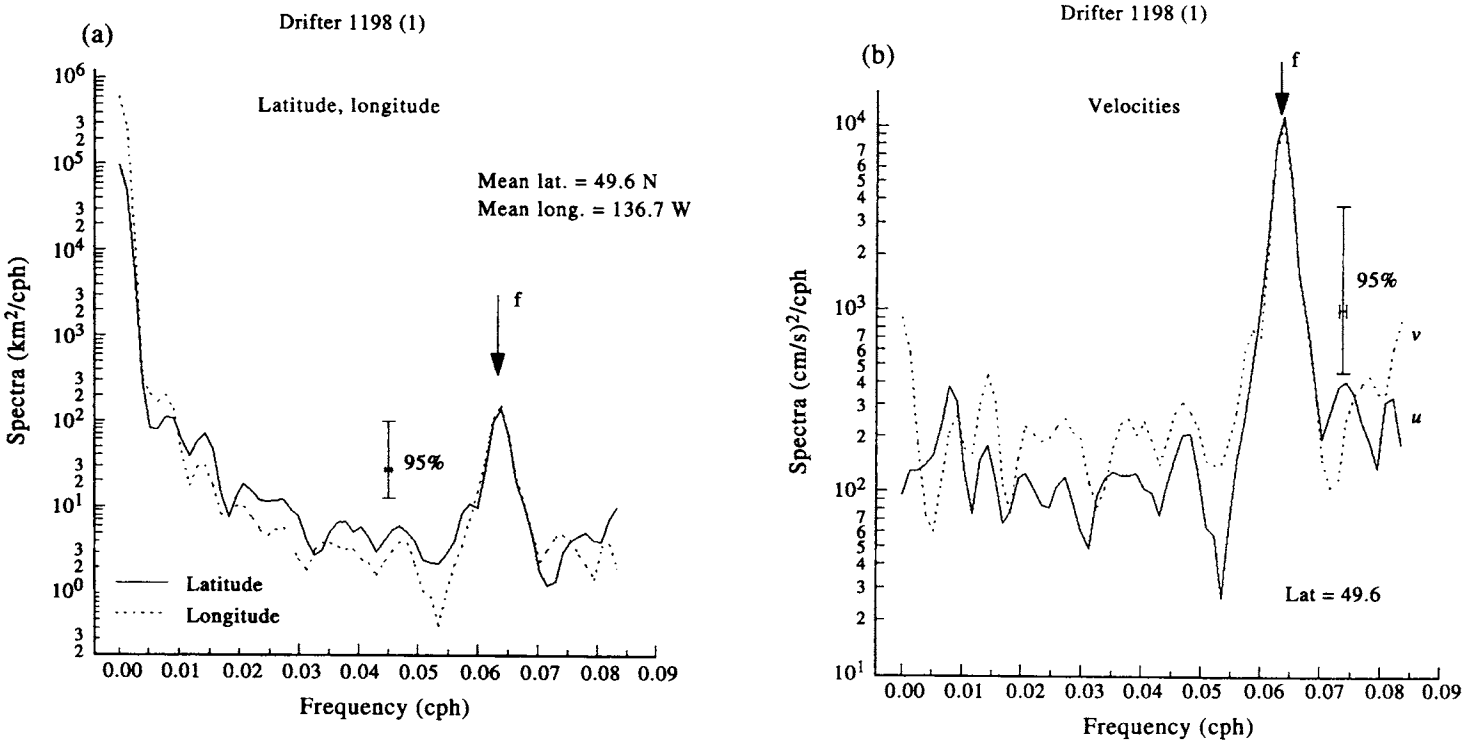


Figure 5.6.28. Effect of a first-difference (high-pass) filter on resulting spectra. (a) Spectra of longitude (Δx) and latitude (Δy) displacements of a satellite-tracked drifter launched in the northeast Pacific in September 1990 (Δt = 3 h; duration T = 90 days); (b) as with (a) but for the zonal ($u = \Delta x/\Delta t$) and meridional velocity ($v = \Delta y/\Delta t$). Mean position of the drifter was 49.6°N, 136.7°W. f denotes the mean inertial frequency; vertical line is the 95% confidence interval.

least-squares spectral analysis method for unevenly spaced time series. The Lomb method described by Press *et al.* (1992) evaluates data, and associated sines and cosines, at the times t_n that the data are measured. For the N data values $x(t_n) = x_n$, $i = 1, \ldots, N$, the Lomb-normalized periodogram is defined as

$$P(\omega) = \frac{1}{2\sigma^2}\left\{\frac{\left[\sum_{n=1}^{N}(x_n - \bar{x})\cos[\omega(t_n - \tau)]\right]^2}{\sum_{n=1}^{N}\cos^2[\omega(t_n - \tau)]} + \frac{\left[\sum_{n=1}^{N}(x_n - \bar{x})\sin[\omega(t_n - \tau)]\right]^2}{\sum_{n=1}^{N}\sin^2[\omega(t_n - \tau)]}\right\} \tag{5.6.102}$$

where as usual

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2 \tag{5.6.103}$$

are the mean and standard deviation of the time series, and the time offset, τ, is defined by

$$\tan(2\omega\tau) = \frac{\sum_{n=1}^{N}\sin(2\omega t_n)}{\sum_{n=1}^{N}\cos(2\omega t_n)} \tag{5.6.104}$$

The offset τ renders equation (5.6.102) identical to the equation we would derive if we attempted to estimate the harmonic content of a data set at frequency ω using the linear least-squares model

$$x(t) = A\cos\omega t + B\sin\omega t \tag{5.6.105}$$

In fact, Vaníček's founding paper on the technique refers to it as a least-squares spectral analysis method. The method, which gives superior results to FFT methods, weights the data on a per point basis rather than on a time-interval basis. By not using weights that span a constant time interval, the method reduces errors introduced by unevenly sampled data. For further details on the Lomb periodogram, including the introduction of significance testing of spectral peaks, the reader is referred to Press *et al.* (1992; pp. 569–577).

5.6.11 General spectral bandwidth and Q of the system

Once the power spectral density, $S(\omega)$, has been computed, the general spectral bandwidth BW may be determined from the three moments, m_k, of the spectra

$$\begin{aligned} m_k &= \int_0^\infty \omega^k S(\omega)\,\mathrm{d}\omega, \quad k = 0, 1, 2 \\ &\approx \sum_{i=0}^{N/2} \omega_i^k S(\omega_i)\Delta\omega \end{aligned} \tag{5.6.106}$$

where $N/2$ is the number of spectral estimates and $\Delta\omega$ is the frequency resolution of the spectral estimates (cf. Masson, 1996). In particular

$$BW = (m_2 m_0 / m_1^2 - 1)^{1/2} \tag{5.6.107}$$

The bandwidth, $\Delta\omega_{BW}$, of a particular spectral peak within an oscillatory system can be used to estimate the dissipation of the system at the peak (resonant) frequency, ω_r. Specifically, the "Q" or *Quality* factor of the system measures the amount of energy stored in a linear oscillator compared to the amount of energy lost per cycle through frictional dissipation. The Q-factor characterizes the sharpness of the resonant frequency and is commonly used as a direct measure of tidal dissipation in the ocean. Suppose that the energy of a simple linear system passes through a maximum at resonance frequency and that the energy of the system falls to 50% of its maximum value at frequencies $\omega \approx \omega_r \pm \Delta\omega_{BW}/2$. The Q of the system is then given by

$$Q = \omega_r / \Delta\omega_{BW} \tag{5.6.108}$$

For example, Wunsch (1972) finds $Q \approx 3.3$ for an apparent resonant period of 14.8 h for the North Atlantic Ocean while Garrett and Munk (1971) obtain an global-wide lower bound of 25 for normal modes near the semidiurnal frequency.

5.6.12 Summary of the standard spectral analysis approach

In summary, PSD estimates for time series $y(t)$ can be obtained as follows using the standard autocorrelation and periodogram approaches:

(1) Remove the mean and trend from the time series. If block averaging is to be used to improve the statistical reliability of the spectral estimates (i.e. to increase the number of degrees of freedom), divide the data series into M sequential blocks of N' data values each, where $N' = N/M$ (see Section 5.6.7).

(2) To partially reduce end effects (Gibbs' phenomenon) or to increase the series length to a power of two for FFT analysis, pad the data with $K \leq N$ zeros. Also pad the record with zeros if you wish to increase the frequency resolution or center spectral estimates in specific frequency bands. To further reduce end effects and side-lobe leakage, taper the time series using a Hanning (raised cosine) window, Kaiser–Bessel window, or other appropriate window (see Section 5.6.6).

(3) Compute the Fourier transforms, $Y(f_k), k = 0, 1, 2, \ldots, N-1$, for the time series (for convenience, we have taken $K = 0$). For block-segmented data, calculate the Fourier transforms $Y_m(f_k)$ for each of the M blocks ($m = 1, \ldots, M$) where $k = 0, 1, \ldots, N'-1$ and $N' < N$. To reduce the variance associated with the tapering in step 2, the transforms can be computed for overlapping segments.

(4) Re-scale the spectra to account for the loss of "energy" during application of the window. That is, adjust the scale factor of $Y(f_k)$ (or $Y_m(f_k)$ in the case of smaller block size partitioning) to account for the reduction in spectral energy due to the tapering in step 2. For the Hanning window, multiply the amplitudes of the Fourier transforms by $\sqrt{(8/3)}$. The rescaling factors for other windows are listed in the right-hand column of Table 5.6.4.

(5) Compute the raw power spectral density for the time series (or for each block) where for the two-sided spectral density estimates

$$S_{yy}(f_k) = \frac{1}{N\Delta t}[Y^*(f_k)Y(f_k)], \quad k = 0, 1, 2, ..., N-1$$

(no block averaging)

$$S_{yy}(f_k; m) = \frac{1}{N\Delta t}[Y_m^*(f_k)Y_m(f_k)], \quad k = 0, 1, 2, ..., N'-1 \tag{5.6.109a}$$

(block averaging)

and for the one-sided spectral density estimates

$$G_{yy}(f_k) = \frac{2}{N\Delta t}[Y^*(f_k)Y(f_k)], \quad k = 0, 1, 2, ..., N/2$$

(no block averaging)

$$G_{yy}(f_k; m) = \frac{2}{N\Delta t}[Y_m^*(f_k)Y_m(f_k)], \quad k = 0, 1, 2, ..., N'/2 \tag{5.6.109b}$$

(block averaging)

(6) In the case of the block-segmented data, average the raw spectral density estimates from the m blocks of data, frequency-band by frequency-band, to obtain the smoothed periodogram for $S_{yy}(f_k)$ or $G_{yy}(f_k)$. Remember, the trade-off for increased smoothing (more degrees of freedom) is a decrease in frequency resolution.

(7) Incorporate 80, 90, and/or 95% confidence limits in spectral plots to indicate the statistical reliability of spectral peaks. Most authors use the 95% confidence intervals.

We can illustrate some of the points in the above summary using the log–log spectra of sea-level oscillations (Figure 5.6.29) recorded over 14 days (20,160 min) in 1991 at Malokurilsk Bay on the west coast of Shikotan Island in the western Pacific. The main spectral peak is centered at a period of 18.6 min and corresponds to a wind-generated seiche amplitude of about 25 cm (Rabinovich and Levyant, 1992). All spectra have been obtained using segmented versions of the 14-day time series. Each time-series segment has been smoothed using a Kaiser–Bessel window with 50% overlap between segments and each segment has been treated as an independent time series. An FFT algorithm was used to calculate the spectrum for each segment. The smoothest spectrum (Figure 5.6.29a) is based on block averaged spectral estimates from roughly 157 overlapping segments (~20,160 min/128 min), the moderately smooth spectrum (Figure 5.6.29b) from the average of 39 overlapping segments, and the noisiest spectrum (Figure 5.6.29c) from the average of 10 overlapping segments. Taking into account the 50% overlap between segments and the fact that there are two degrees of freedom (DOF) per raw spectral estimate, there are 628 (= 157 × 4), 154, and 36 DOF for the three spectra, respectively. The smoothed spectrum in Figure 5.6.29(d) is derived using a slightly different approach. Although the segment lengths are the same as those in Figure 5.6.29(c) (i.e. 2048 min), the number of DOF is increased with increasing frequency, ω. In this sliding scale, the lowest frequency range uses 36 DOF (as with Figure 5.6.29c), the next frequency band averages together the spectra for

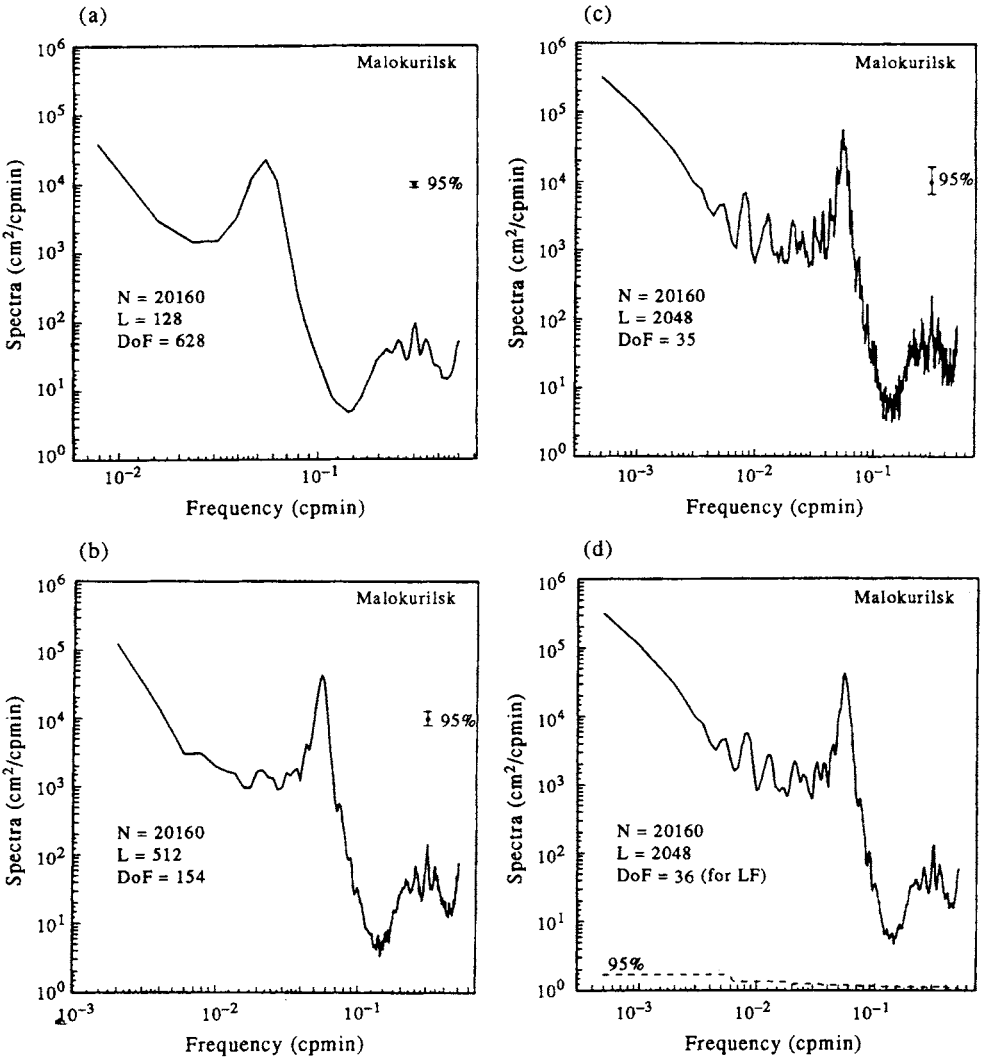


Figure 5.6.29. Spectra of sea-level oscillations recorded by a bottom-pressure gauge in Malokurilsk Bay on the west coast of Shikotan Island. Time-series length $T = N\Delta t$, *where* $N = 20{,}160$ *and* $\Delta t = 1$ *min. Segment lengths are* $T_s = M\Delta t$, $M \ll N$. *Each time-series segment has been smoothed with a Kaiser–Bessel window with 50% overlap between segments. Block averaging has been used to smooth the spectral estimates. (a) Highly smoothed spectrum with* $M = 128$ (2^7), *DOF = 628; (b) moderately smoothed spectrum with* $M = 512$ (2^9), *DOF = 154; (c) weakly smoothed spectrum with* $M = 2048$ (2^{11}), *DOF = 36; (d) same as (c) except that DOF = 36 applies to the lowest frequency range only. For* $f \geq 6 \times 10^{-2}$ *cycles/min, the number of spectral estimates averaged together increases as* 3×36, 5×36, *and* 7×36, *for each of the next three frequency ranges. (Courtesy of A. Rabinovich.)*

three adjacent frequencies to give 108 DOF, the next averages together the spectra for five adjacent frequencies to give 180 DOF, and so on.

As indicated by Figure 5.6.29, increasing the number of frequency bands averaged in each spectral estimate enhances the overall smoothness of the spectrum and improves the statistical reliability for specific spectral peaks. The number of DOFs increases and the confidence interval narrows. The penalty we pay for improved

statistical confidence is reduced resolution of the spectral peaks. As in Figure 5.6.29(a), too much smoothing diminishes our ability to specify the frequency of spectral peaks and washes out peaks linked to some of the weaker seiches. Because each time-series segment is so short, we also lose definition at the low-frequency end of the spectrum. As indicated by Figure 5.6.29(c), too little smoothing leads to a noisy spectrum for which few spectral peaks are associated with any physical processes. The sliding DOF scale in Figure 5.6.29(d) is a useful compromise.

Covariance function: Since the covariance function, $C_{yy}(\tau)$, and the autospectrum are Fourier transform pairs, the above analysis can be used to obtain a smoothed or unsmoothed estimate of the covariance function. To do this, first calculate the Fourier transform $Y(f)$ of the time series, and determine the product $S_{yy}(f) = N^{-1}\Delta t[Y^*(f)Y(f)]$. Then take the inverse Fourier transform (IFT) of the autospectrum, $S_{yy}(f)$, to obtain the covariance function, $C_{yy}(\tau)$. If the spectrum is unsmoothed prior to the IFT (or IFFT if the FFT was used), we obtain the raw covariance function. If, on the other hand, the autospectrum is smoothed prior to the above integral using one of the spectral windows, such as the Hanning window, the covariance function also will be a smoothed function.

A word of caution: Although everyone agrees on the basic formulation for the discrete Fourier transform (DFT) and the inverse discrete Fourier transform (IDFT), there are several ways to normalize the relations using the number of records, N. In our definitions, (5.6.10) and (5.6.12), N appears in the denominator of the inverse discrete Fourier transform. Some authors normalize using $1/N$ in the DFT only while others insist on symmetry by using $1/\sqrt{N}$ in both DFT and its inverse. User alert: When using "canned" programs to obtain DFTs and IDFTs, ensure that you know how the transforms are defined and adapt your analysis to fit the appropriate processing routines.

5.7 SPECTRAL ANALYSIS (PARAMETRIC METHODS)

If the analytical model for a time series was known exactly, a sensible spectral estimation method would be to fit the model spectrum to the observed spectrum and determine any unknown parameters. In general, however, oceanic variability is too complex to admit simple analytical models and parametric spectral estimates over the full frequency range of the data series. In addition, the imposition of an overly simplified spectral model could seriously degrade any estimation. On the other hand, it is reasonable that relatively simple spectral models might adequately reflect the system dynamics over limited frequency bands. Under some very general conditions, any stationary series can be represented in closed form by a statistical model in which the corresponding spectrum is a rational function of frequency (i.e. a ratio of two polynomials in ω).

If the time series under investigation is long relative to the time scales of interest, and if the spectrum is not overly complicated and does not have too large a dynamic range, the simple smoothed periodogram technique will probably yield adequate results. At a minimum, it will identify the major features in the spectrum. For shorter time series or in studies of fine spectral structure, other techniques may be more applicable. One such spectral analysis technique was developed by Burg (1967, 1972)

who showed that it was possible to obtain the power spectrum by requiring the spectral estimate to be the most random or have the maximum entropy of any power spectrum which is consistent with the measured data. This leads to a spectral estimate with a high frequency resolution since the method uses the available lags in the autocovariance function without modification and makes a nonzero estimate (prediction) of the autocorrelation function beyond those which are routinely calculated from the data. Because the spectral values are computed using a maximum entropy condition, the resulting spectral estimates are not accurate in terms of spectral amplitude.

The most popular of the "modern" parametric techniques is the *Autoregressive power spectral density* (AR PSD) model whose origins are in economic time series forecasting and statistical estimation. Autoregressive estimation was introduced to the earth sciences in the 1960s where it was originally applied to geophysical time-series data under the name *maximum entropy method* (MEM). The duality between AR and MEM estimation has been thoroughly explored by Ulrych and Bishop (1975). Autoregressive spectral estimation is attractive because it has superior frequency resolution compared to conventional FFT techniques. As an example of the frequency resolution capability, consider the 14-year time series of average monthly air temperature for New York city (Figure 5.7.1a). The unsmoothed periodogram and three smoothed periodograms reveal a broad spectral peak centered at a period of one year (Figure 5.7.1b). This compares to the much sharper annual peak obtained via AR estimation (Figure 5.7.1c). The results reveal another important difference between the two methods. With the nonparametric periodogram approach, we can determine confidence limits for the spectral peaks while for the parametric method the significance of the peaks is unknown. For example, the maximum entropy method is good for finding the location of spectral peaks but is not reliable for computing the correct spectral energy at those peaks. (The periodogram smoothing in Figure 5.7.1(b) was performed using a Parzen window with truncation values $N = 16$, 32, and 64; the weights for these windows are $w(n) = 1.0 - |2n/N|^2$, with $0 \leq |n| \leq \frac{1}{2}N$.)

In general, autoregressive and maximum entropy PSD estimation are not as widely used in oceanography as traditional spectral analysis methods. The former find their greatest application in analytical climate modeling and in wavenumber spectral estimation. Modern parametric techniques are good so long as the model is good. On the other hand, if the model is false, the resulting spectrum estimate can be highly misleading. It follows that if you have no reason for believing a specific model you are better served using a nonparametric model. For this reason, we limit our presentation to the essential elements of the two methods. The reader is directed to Marple (1987) for a thorough discussion of the topic, including an introduction to Fourier transform methods of spectral analysis.

5.7.1 Some basic concepts

Many deterministic and stochastic discrete-time series processes encountered in oceanography are closely approximated by a rational transfer model in which the input sequence $\{x_n\}$ and the output sequence $\{y_n\}$, which is meant to model the input data, are related by the linear difference relation

$$y_n = \sum_{k=0}^{q} b_k x_{n-k} - \sum_{m=1}^{p} a_m y_{n-m} \tag{5.7.1}$$

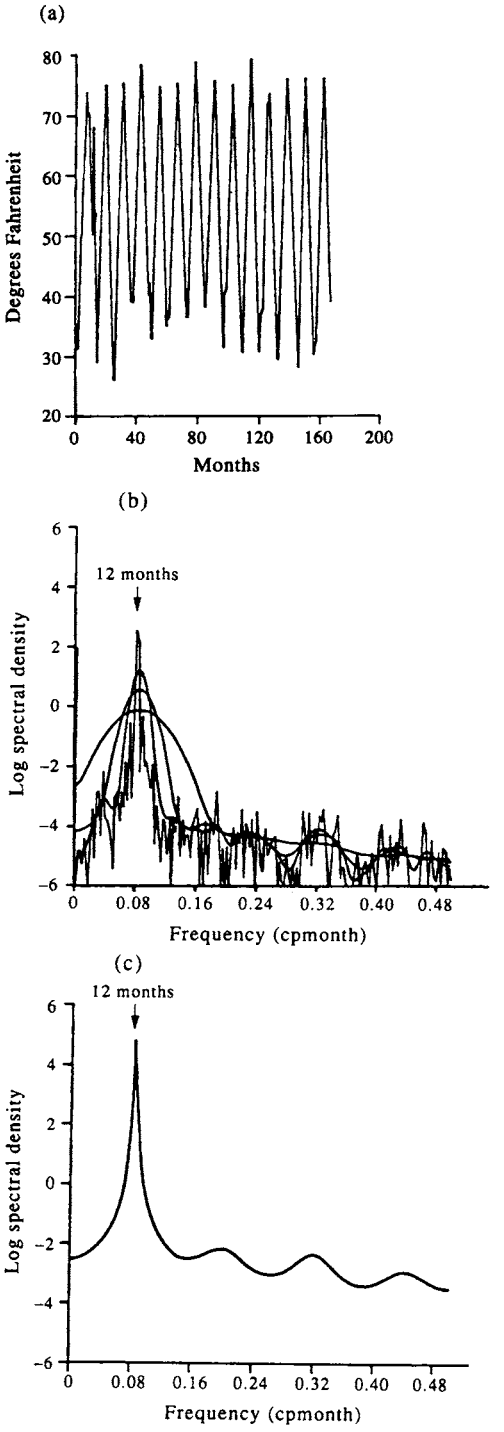


Figure 5.7.1. (a) Time series of monthly average air temperature for New York city (1946–1959); (b) the unsmoothed (raw) periodogram and three smoothed periodograms for Parzen windows with truncation lengths of 16, 32, and 64 months; and (c) an autoregressive spectral estimate of (a) showing the sharp peak at 12-months period. (From Pagano, 1978.)

Here, y_n is shorthand notation for $y(n\Delta t)$, also written as $y(n)$. In its most general form, the linear model (5.7.1) is termed an *autoregressive moving average* (ARMA) model. The power spectral density (PSD) of the ARMA output process is

$$P_{ARMA}(f) = \sigma^2 \Delta t |A(f)/B(f)|^2 \tag{5.7.2}$$

where σ^2 is the variance of the applied white-noise driving mechanism and $\sigma^2 \Delta t$ is the PSD of the noise for the Nyquist interval $-1/(2\Delta t) \leq f \leq 1/(2\Delta t)$. Here

$$A(f) = \alpha[\exp(i2\pi f \Delta t)], \quad B(f) = \beta[\exp(i2\pi f \Delta t)] \tag{5.7.3}$$

where the coefficients α, β are defined in terms of the *z-transform*, $X(z)$, of the variable $z = \exp(i2\pi f \Delta t)[= \exp(i2\pi nk/N)$ in discrete form] where $k, n = 0, 1, \ldots, N-1$

$$X(z) = \sum_{n=0}^{N-1} x_n z^{-n} \tag{5.7.4}$$

which maps a real-valued sequence into a complex plane. Note that equation (5.7.4) is defined through negative powers of z, the convention used in electrical engineering. Geophysicists expand in positive powers of z (z^{+n}) but define $z = \exp(-iz\pi f \Delta t)$ so the results are the same. The z-transform of the autoregressive branch is

$$\alpha(z) = \sum_n a_n z^{-n} \tag{5.7.5a}$$

while that of the moving average branch is

$$\beta(z) = \sum_n b_n z^{-n} \tag{5.7.5b}$$

Specification of the parameters $\{a_k\}$, termed the autoregressive coefficients, the parameters $\{b_k\}$, termed the moving-average coefficients, and the variance σ^2 is equivalent to specifying the spectrum of the process $\{y_n\}$. Without loss of generality, one can assume $a_o = 1$ and $b_o = 1$ since any gain of the system (5.7.1) can be incorporated into σ^2. If all the $\{a_k\}$ terms except $a_o = 1$ vanish then

$$y_n = \sum_{k=0}^{q} b_k x_{n-k} \tag{5.7.6}$$

and the process is simply a moving average of order q, and

$$P_{MA}(f) = \sigma^2 \Delta t |A(f)|^2 \tag{5.7.7}$$

This model is sometimes called an *all-zero model* since spectral peaks and valleys are formed through zeros of the function $A(f)$. If all the $\{b_k\}$ terms except $b_o = 1$ vanish then

$$y_n = \sum_{m=1}^{p} a_m x_{n-m} + \varepsilon_n \tag{5.7.8}$$

and the process is strictly an autoregressive model of order p. The process is called AR in the sense that the sequence x_n is a linear regression on itself with ε_n representing

the error. With this model, the present value y_n is expressed as a weighted sum of past values plus a noise term. The PSD is

$$P_{AR}(f) = \frac{\sigma^2 \Delta t}{|B(f)|^2} \tag{5.7.9}$$

In the engineering literature, this model is sometimes called an *all-pole model* since narrow spectral peaks can be sharply delineated through zeros in the denominator.

5.7.2 Autoregressive power spectral estimation

The discrete form of an autoregressive model $y(t)$ of order p is represented by the relationship

$$y(n) = a_1 y(n-1) + a_2 y(n-2) + \ldots + a_p y(n-p) + \varepsilon(n) \tag{5.7.10}$$

where time $t = n\Delta t$, the a_k $(k = 1, \ldots, p)$ are constant coefficients, and $\varepsilon(t)$ is a white-noise series (usually called the innovation of the AR process) with zero mean and variance σ^2. Another interpretation of the AR process links $y(t)$ with a value that is predicted from the previous $p-1$ values of the process with a prediction error equal to $\varepsilon(t)$. Thus, the a_k $(k = 1, \ldots, p)$ represent a p-point prediction filter. If $Y(z)$ is the z-transform of $y(n)$ then

$$Y(z) = \sum_{n=0}^{p} y(n) z^n \tag{5.7.11}$$

and

$$Y(z) - Y(z)(a_1 z + a_2 z^2 + \ldots + a_p z^p) = D(z) \tag{5.7.12}$$

so that

$$|Y(z)|^2 = \frac{|D(z)|^2}{|1 - a_1 z - a_2 z^2 \ldots - a_p z^p|^2} \tag{5.7.13}$$

Substituting $z = \exp(-\mathrm{i}2\pi f \Delta t)$ we obtain half of the true power spectrum. If the autoregression is a reasonable model for the data, then the autoregressive power spectral density estimate based on (5.7.9) is

$$P_{AR}(f) = \frac{\sigma^2 \Delta t}{\left|1 + \sum_{k=1}^{p} a_k \exp(-\mathrm{i}2\pi f k \Delta t)\right|^2} \tag{5.7.14}$$

To find the PSD we need only estimate three things: (1) the autoregressive parameters $\{a_1, a_2, \ldots, a_p\}$; (2) the variance, σ^2, of the white-noise process that is assumed to be driving the system; and (3) the order, p, of the process. The limitations of the AR model are the degrading effect of observational noise, spurious peaks, and some anomalous effects which occur when the data are dominated by sinusoidal components. Unlike conventional Fourier spectral estimates, the peak amplitudes in AR spectral estimates are not linearly proportional to the power when the input process consists of sinusoids in

noise. For high signal-to-noise ratios, the peak is proportional to the square of the power with the area under the peak proportional to power.

5.7.2.1 Autoregressive parameter estimation

Yule–Walker equations: If the autocorrelation function, $R_{yy}(k)$, is known exactly, we can find the $\{a_k\}$ by the Yule–Walker equations. This method relates the AR parameters to the known (or estimated) autocorrelation function of $y(n)$

$$R_{yy}(k) = \frac{1}{N}\sum_{n=1}^{N-k}[[x(n+k)-\bar{x}][x(n)-\bar{x}]; \quad \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x(n) \tag{5.7.15}$$

There are other methods of estimating R_{yy} but this estimator has the attractive property that its mean-squared error is generally smaller than that of other estimators (Jenkins and Watts, 1968). Since it is generally assumed that the mean $\bar{x}$ has been removed from the data, the autocovariance and autocorrelation functions are identical. To get the AR parameters, one need only choose p equations from the Yule–Walker equations for $k > 0$, solve for $\{a_1, a_2, \ldots, a_p\}$, and then find σ^2 from (2.39) for $k = 0$. The matrix equation to find the a_is and σ^2 is

$$\begin{vmatrix} R_{yy}(0) & R_{yy}(-1) & \ldots & R_{yy}(-p) \\ R_{yy}(1) & R_{yy}(0) & \ldots & R_{yy}[-(p-1)] \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ R_{yy}(p) & R_{yy}(p-1) & \ldots & R_{yy}(0) \end{vmatrix} \begin{vmatrix} 1 \\ a_1 \\ \ldots \\ \ldots \\ a_p \end{vmatrix} = \begin{vmatrix} \sigma^2 \\ 0 \\ \ldots \\ \ldots \\ 0 \end{vmatrix} \tag{5.7.16}$$

Thus, to determine the AR parameters and the variance σ^2 one must solve (5.7.16) using the $p + 1$ autocorrelation lags, $R_{yy}(0), \ldots, R_{yy}(p)$, where $R_{yy}(-k) = R_{yy}^*(k)$.

Solutions to the Yule–Walker matrix equation can be found via the computationally efficient Levinson–Durbin algorithm which proceeds recursively to compute the parameter sets $\{a_{11}, \sigma_1^2\}, \{a_{21}, a_{22}, \sigma_2^2\}, \ldots, \{a_{p1}, a_{p2}, \ldots, a_{pp}, \sigma_p^2\}$. The final set at order p (the first subscript) is the desired solution. The algorithm requires p^2 operations as opposed to the $O(p^3)$ operations of Gaussian elimination. More specifically, the recursion algorithm gives

$$a_{11} = \frac{-R_{yy}(1)}{R_{yy}(0)} \tag{5.7.17a}$$

$$\sigma_1^2 = (1 - |a_{11}|^2)R_{yy}(0) \tag{5.7.17b}$$

with the recursion for $k = 2, 3, \ldots, p$ given by

$$a_{kk} = \frac{-1}{\sigma_1^2}\left[R_{yy}(k) + \sum_{j=1}^{k-1} a_{k-1,j} R_{yy}^{(k-j)}\right] \tag{5.7.18a}$$

$$a_{ki} = -a_{k-1,i} + a_{kk}(a_{k-1,k-i})^* \tag{5.7.18b}$$

$$\sigma_k^2 = (1 - |a_{kk}|^2)\sigma_{k-1}^2 \tag{5.7.18c}$$

Burg algorithm: Box and Jenkins (1970) point out that the Yule–Walker estimates of

the AR coefficients are very sensitive to rounding errors, particularly when the AR process is close to becoming nonstationary. The assumption that $y(k) = 0$, for $|k| > p$ leads to a discontinuity in the autocorrelation function and a smearing of the estimated PSD. For this reason, the most popular method for determining the AR parameters (prediction error filter coefficients) is the Burg algorithm. This algorithm works directly on the data rather than on the autocorrelation function and is subject to the Levinson recursion (5.7.18b). As an illustration of the differences in the YW and the Burg estimates, the respective values of a_{11} for the series $y(t_k) = y(k)$ are

$$\begin{aligned} a_{11} &= \frac{\sum_{k=2}^{p} y(k)y(k-1)}{\sum_{k=1}^{p} y(k)^2}, \text{ for the Yule-Walker estimate} \\ a_{11} &= \frac{\sum_{k=2}^{p} y(k)y(k-1)}{\frac{1}{2}x_1^2 + \sum_{k=1}^{p} y(k)^2 + \frac{1}{2}x_p^2}, \text{ for the Burg estimate} \end{aligned} \tag{5.7.19}$$

Detailed formulation of the Burg algorithm is provided by Kay and Marple (1981; p. 1392). Again, there are limitations to the Burg algorithm, including spectral line splitting and biases in the frequency estimate due to contamination by rounding errors. Spectral line splitting occurs when the spectral estimate exhibits two closely spaced peaks, falsely indicating a second sinusoid in the data.

Least squares estimators: Several least squares estimation procedures exist that operate directly on the data to yield improved AR parameter estimates and spectra than the Yule–Walker or Burg approaches. The two most common methods use forward linear prediction for the estimate, while a second employs a combination of forward and backward linear prediction. Ulrych and Bishop (1975) and Nuttall (1976) independently suggested this least squares procedure for forward and backward prediction in which the Levison recursion constraint imposed by Burg is removed. The least squares algorithm is almost as computationally efficient as the Burg algorithm requiring about 20 more computations. The improvement by the LS approach over the Burg algorithm is well worth the added computation time. Improvements include less bias in the frequency estimates, and absence of observed spectral line splitting for short sample sinusoidal data.

Barrodale and Erickson (1978) provide a FORTRAN program for an "optimal" least-squares solution to the linear prediction problem. The algorithm solves the underlying least-squares problem directly without forcing a Toeplitz structure on the model. Their algorithm can be used to determine the parameters of the AR model associated with the maximum entropy method and for estimating the order of the model to be used. As illustrated by the spectra in Figure 5.7.2, this approach leads to a more accurate frequency resolution for short sample harmonic processes. In this case, the test data were formed by summing 0.03 and 0.2 Hz sine waves generated in single precision and sampled 10 times per second. The reader is also referred to Kay and Marple (1981; p. 1393) for additional details.

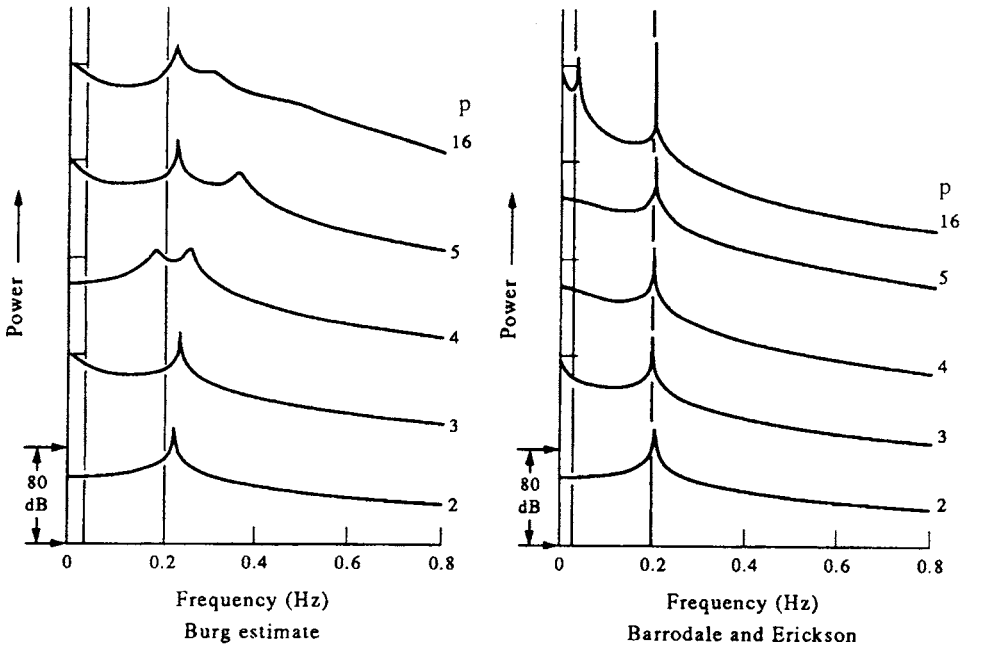


Figure 5.7.2. Maximum entropy method spectra obtained using (a) the Burg and (b) the Barrodale and Erickson algorithms. Signal consists of a combined 0.2 and 0.03 Hz (cps) sine wave. Spectra are plotted for increasing numbers of coefficients, p. (From Barrodale and Erickson, 1978.)

5.7.2.2 Order of the autoregressive process

The order p of the autoregressive filter is generally not known *a priori* and is acknowledged as one of the most difficult tasks in time series modeling by parametric methods. The choice is to postulate several model orders then compute some error criterion that indicates which model order to pick. Too low a guess for the model order results in a highly smoothed spectral estimate. Too high an order introduces spurious detail into the spectrum. One intuitive approach would be to construct AR models with increasing order until the computed prediction error power σ_k^2 reaches a minimum. Thus, if a process is actually an AR process of order p, then $a_{p+1,\,k} = a_{pk}$ for $k = 1, 2, \ldots, p$. The point at which a_{pk} does not change would appear to be a good indicator of the correct model order. Unfortunately, both the Yule–Walker equations and Burg algorithm involve prediction error powers

$$\sigma_k^2 = \sigma_{k-1}^2[1 - |a_{kk}|^2] \tag{5.7.20a}$$

that decrease monotonically with increasing order p, so that as long as $|a_{kk}|^2$ is nonzero (it must be ≤ 1) the prediction error power decreases. Thus, the prediction error power is not sufficient to indicate when to terminate the search. Alternative approaches (Kay and Marple, 1981) have been proposed by Akaike (termed the final prediction error, FPE, and the Akaike information criterion, AIC), and by Parzen (termed the criterion autoregressive transfer function). The Akaike information criterion determines the model order by minimizing an information theoretic function. If the process has Gaussian statistics, the AIC is

Summary of algorithms

Method	Model applied	Advantages	Disadvantages
Periodogram method using FFT or direct Fourier transform	Sum of harmonics (sines and cosines). No specific model needed.	1. Uses harmonic least squares fit to the data; 2. output $S(f)$ directly proportional to power; 3. most computationally efficient; 4. well-established methodology; 5. confidence intervals easily computed; 6. integral of $S(f)$ over frequency band Δf is equal to the variance of the signal in that band. 7. easily generalized to cross-spectra and rotary spectra analysis.	1. Frequency resolution $\Delta f \approx 1/T$ dependent only on record length, T; 2. poor performance for short data records; 3. side-lobe leakage distorts spectra if appropriate windowing not done; windowing reduces frequency resolution, Δf; 4. must average spectral estimates to improve statistical reliability.
Autoregressive, Yule–Walker algorithm.	Autoregressive (all-pole) process. Specific model.	1. Improved spectral resolution over Fourier transform methods; 2. sharp spectral peaks; 3. no side-lobe leakage problems; 4. minimum phase (stable) linear prediction filter guaranteed if biased lag estimates computed; 5. related to linear prediction analysis and adaptive filtering.	1. Model order, p, must be specified; 2. spectral line splitting occurs; 3. implied windowing distorts spectra; 4. confidence intervals not readily computed.
Autoregressive, Burg algorithm.	Autoregressive (all-pole) process. Specific model.	1. Improved resolution over Fourier transform methods. Uses a constrained recursive least squares approach 2. no side-lobe leakage problems; 3. high resolution for low noise signals; 4. good spectral fidelity for short data series; 5. no windowing implied; 6. Stable linear prediction filter guaranteed.	1. Model order, p, must be specified; 2. spectral line splitting can occur; 3. confidence intervals not readily computed.
Autoregressive, least-squares method.	Autoregressive (all-pole) process. Specific model.	1. Sharper spectra than for other AR methods 2. no side-lobes; 3. good spectral fidelity for short data series; 4. no windowing; 5. no line splitting; 6. uses exact recursive least squares solution with no constraint.	1. Model order must be specified; 2. stable linear prediction filter not guaranteed, though stable filter results in most cases.

$$\mathrm{AIC}(p) = \ln(\sigma_p^2) + 2(p+1)/N \tag{5.720b}$$

where σ_p^2 is the prediction error power and N is the number of data samples. The second term represents the penalty for the use of extra autoregressive coefficients that do not result in a substantial reduction in the prediction error power. The order p is the one that minimizes the AIC.

5.7.2.3 Maximum entropy method (MEM)

The only constraint on the AR method is that the data yield the known autocorrelation function $R_{yy}(k)$ for the interval $0 < k < p$. The assumption that $y(k) = 0$, for $|k| > p$ leads to a discontinuity in the autocorrelation function and a smearing of the estimated power spectral density. The MEM was designed, independently of autoregressive estimation, to eliminate the distortion of the spectrum caused by the truncated $R_{yy}(k)$. By adding a second constraint to improve the spectral estimation, the method gets away from the problems with the Yule–Walker algorithm. In essence, the MEM is a way of extrapolating the known autocorrelation function to lags $k > p$, which are not known. In words, we assume that $\{R_{yy}(0), \ldots, R_{yy}(p)\}$ are known and find a logical way to extend to lags $\{R_{yy}(p+1), \ldots\}$. As it turns out, the power spectral estimate for the MEM approach is equivalent to the power spectral estimate for the AR process.

In general, there exist an infinite number of possible extrapolations. Burg (1968) argued that preferred extrapolation should do two things: (1) Yield the known R_{yy} for $0 \leq k \leq p$; and (2) generate an extrapolated R_{yy} for $k > p$ that causes the time series to have maximum entropy under the constraint (1). The time series that results is the most random one which adheres to the known R_{yy} for the first $p + 1$ lags. Alternatively, we can say that PSD is the one with whitest noise (flattest spectrum) of all possible spectra for which $\{R_{yy}(0), \ldots, R_{yy}(p)\}$ is known. The reason for choosing the maximum entropy criterion is that it imposes the fewest constraints on the unknown time series by maximizing its randomness thereby causing minimum bias and operator intervention. For a Gaussian process, the entropy per sample is proportional to

$$\int_{-1/2\Delta t}^{1/2\Delta t} \ln[P_y(f)]\,\mathrm{d}f \tag{5.7.21}$$

where $P_y(f)$ is the PSD of y_n. The spectrum is found by maximizing (5.7.21) subject to the constraint that the $p + 1$ known lags satisfy the Wiener–Khinchin relation

$$\int_{-1/2\Delta t}^{1/2\Delta t} P_y(f)\mathrm{e}^{-\mathrm{i}2\pi fn\Delta t}\mathrm{d}f = R_{yy}(n), \quad n = 0, 1, \ldots, p \tag{5.7.22}$$

The solution is found using the Lagrange multiplier technique (see Ulrych and Bishop, 1975) as

$$P_y(f) = \frac{\sigma_p^2 \Delta t}{\left|1 + \sum_{k=1}^{p} a_{pk} \exp(-\mathrm{i}2\pi fk\Delta t)\right|^2} \tag{5.7.23}$$

where $\{a_{p1}, \ldots, a_{pp}\}$ and σ_p^2 are just the order-p predictor parameters and prediction error power, respectively. With knowledge of $\{R_{yy}(0), R_{yy}(1), \ldots, R_{yy}(p)\}$ the power spectral density (PSD) of the maximum entropy method (MEM) is equivalent to the PSD of the autoregressive method. That is, the MEM spectral analysis is equivalent to fitting an AR model to the random process. It is indeed interesting that the representation of a stochastic process by an AR model is that representation that exhibits maximum entropy. The duality of the AR model and MEM has enabled workers to apply the large body of literature on AR time-series analysis to overcome shortcomings of the MEM.

The estimation of the MEM spectral density requires a knowledge of the order of the AR process that we use to model the data. The importance of correctly estimating the order p is illustrated using the AR process $y_n \equiv y(t_n)$ at times $t_n = n\Delta t$

$$y_n = 0.75y_{n-1} - 0.5y_{n-2} + \varepsilon_n \tag{5.7.24}$$

with noise variance $\sigma_\varepsilon^2 = 1$ (Figure 5.7.3a). Here $E[y(t)\varepsilon(t)] = \sigma_\varepsilon^2$, but $E[y(t)\varepsilon'(t)] = 0$ for any other additive noise, ε'. As indicated by Figure 5.7.3(b), which compares the theoretical power of a specified second-order AR process with the power spectral density computed from a realization of this process using $p = 2$ and $p = 11$ (Ulrych and Bishop, 1975), the correct choice of p is vital in obtaining a meaningful estimate of the power spectrum of the process. The peak value and the width of the spectral line of the MEM power spectral density estimate also may have considerable variance in the MEM estimates.

Although the MEM has numerous advantages over traditional nonparametric spectral techniques, especially for short data series, the usefulness of the approach is diminished by the lack of a straightforward criterion for choosing the length (order) of the prediction model. Too short a length results in a highly smoothed spectrum obviating the resolution advantages of the MEM, whereas an excessive length introduces spurious detail into the spectrum.

Confidence intervals: A major shortcoming of MEM is the lack of a mathematically consistent variance estimator (confidence interval) for the spectral density. One approach is to approximate the confidence bounds in the same way that we compute the bounds in traditional multivariate spectral analysis (i.e. using a chi-square variable with ν degrees of freedom) under the assumption that the equivalent number of degrees of freedom is given by $\nu = N/p$, where N the number of data points in the time series and p is the order of the model (Privalsky and Jensen, 1993, 1994). The order p should be chosen on the basis of objective criteria such as Akaike's information criterion, Parzen's criterion and so on (see Lütkepohl, 1985).

5.7.2.4 An autoregressive model of global temperatures

One way to determine the effect of initial conditions and random noise on the global temperature predictions of computer-simulated general circulation models (GCMs) is to obtain a control realization, modify the initial conditions and noise, obtain a second realization and compare results. Since this could take several months of supercomputing time, a more practical approach is to employ a model of the global air temperature series, $T(t)$, derived by Jones (1988) (Figure 5.7.4). If we assume that the sensitivity of GCMs to changing conditions is similar to that of a stationary

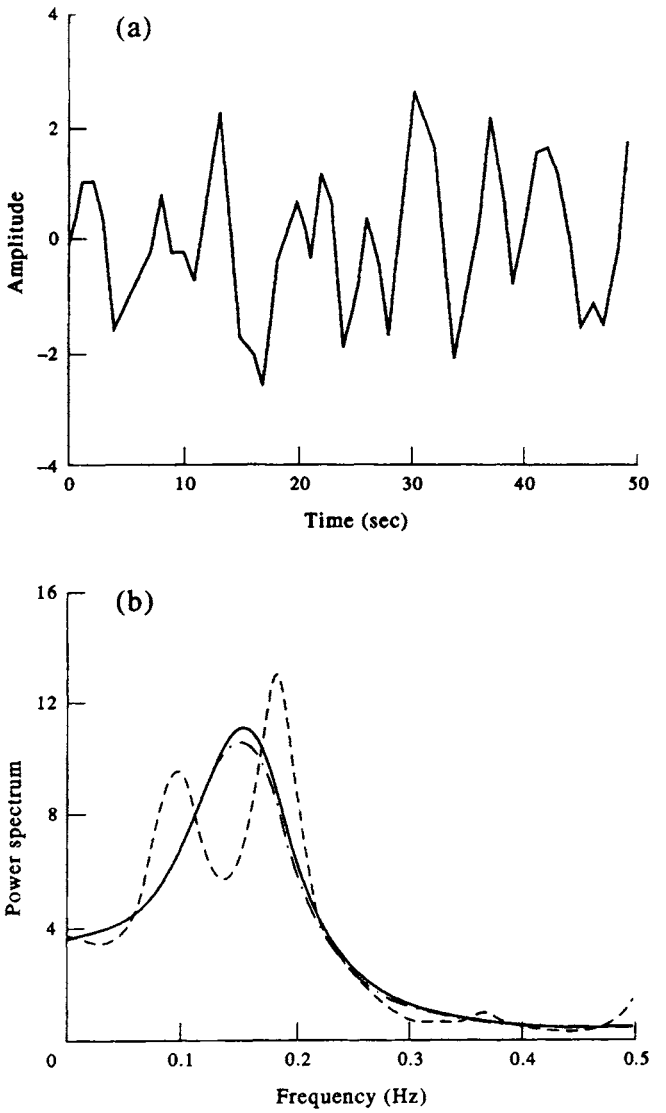


Figure 5.7.3. Maximum entropy spectra. (a) Time series for the second-order AR process $y_n = 0.75y_{n-1} - 0.5y_{n-2} + \varepsilon_n$ *(5.7.24). (b) Spectral computation for the AR process. Solid line: the true power spectrum. Dot–dash line: maximum entropy method (MEM) estimate with 3-point (*$p = 2$*) prediction error filter. Dashed line: MEM estimate with 12-point (*$p = 11$*) error filter. (From Ulrych and Bishop, 1975.)*

autoregressive model, then marked changes in the AR model that result from slight changes in the initial conditions or inherent noise are evidence that GCMs are too sensitive to these parameters to be reliable.

If $Z_n \equiv Z(t_n)$ represents the temperature deviation (departure from the long-term mean) at year t_n, then the maximum likelihood fourth order AR model for the temperature data in Figure 5.7.4 is

$$Z_n = 0.669Z_{n-1} - 0.095Z_{n-2} + 0.104Z_{n-3} + 0.247Z_{n-4} + \varepsilon_n \tag{5.7.25}$$

where $Z_n = T_n - \overline{T}$, and ε_n is an uncorrelated white-noise series with zero mean and variance equal to $0.0115°C^2$ (Tsonis, 1991; Gray and Woodward, 1992). In general, we can state that for any AR process, the initial values will have little effect on forecasts if the sample size is large relative to the order of the process. For this reason, AR processes are often known as "short memory" processes. In the above model, the correlation between $Z(t)$ and $Z(t + m\Delta t)$ is $0.9(0.96)^m$, for values of m greater than about five. For example, the correlation between $Z(t)$ and $Z(t + 30\Delta t)$ is 0.27, while that between $Z(t)$ and $Z(t + 50\Delta t)$ is 0.14. These correlations imply that, even if we started the model with the same initial values $Z_1, \ldots, Z_4$, different realizations of the model would typically have low cross-correlation after 30 years and possess very little similarity beyond 50 years (Figure 5.7.5a). The dissimilarity is associated with the stochastic nature of the noise $\varepsilon(t)$ which quickly decorrelates the present value of the model from its past values. The fact that the two series converge to a similar level near $t = 100$ years is not an indication that they are merging since extending these realizations causes them to depart from one another.

To show the importance of the noise, rather than the initial conditions, Gray and Woodward generated two samples with different starting values but with the same noise sequence. This was intended to mimic a specified set of random conditions driving the weather but having different starting values. As revealed by Figure 5.7.5(b), the realizations begin to merge by year 30, demonstrating their insensitivity to the initial conditions. A further point is that for stationary AR processes, the forecast function is only a function of the sample mean and the last four observations. Since the starting values are independent of the last four observations and small changes in the starting conditions have little effect on the sample mean for a long time series, the forecasts from such a model will be insensitive to changes. In closing their article, Gray and Woodward note that conventional autoregressive moving average (ARMA) modeling methodology indicates that the temperature time series should first be differentiated. Application of a variety of techniques suggests an order 10 (AR(10)) model as the "optimum" model for the differentiated data which gives rise to an AR(11) model for the original time series, not an AR(4) model used in the analysis.

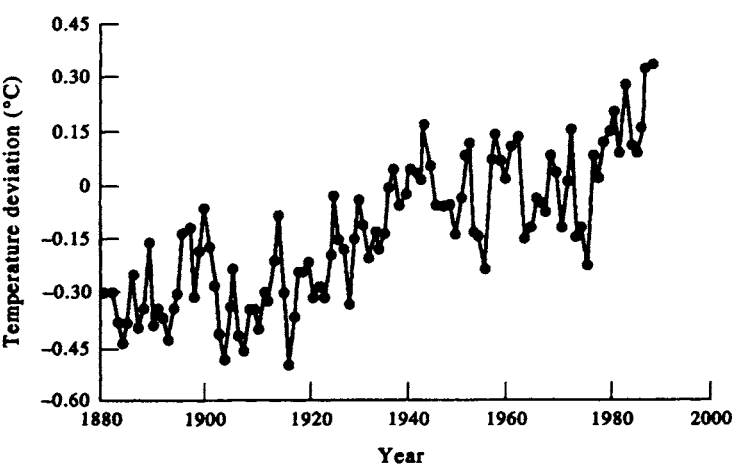


Figure 5.7.4. The annual global mean air temperatures from 1881 to 1988 as deviations (°C) from the 1951–1970 average. (From Gray and Woodward, 1992.)

Lastly, Tsonis (1992) replies that it is not appropriate to change the noise of the signal without also changing the initial conditions.

5.7.3 Maximum likelihood spectral estimation

As first demonstrated by Capon (1969), spectra can be defined using the maximum likelihood procedure. Instead of using a fixed window to operate on the autocorrelation function, the window shape is changed as a function of wavenumber or frequency. The window is designed to reject all frequency components in an optimal way, except for the one frequency component which is desired.

Rather than go through the details of defining the procedure for the maximum likelihood spectrum, we offer here comparisons between the traditional method (in this case, represented by a spectrum computed using a Bartlett window), a maximum likelihood spectrum, and a spectrum computing using the maximum entropy

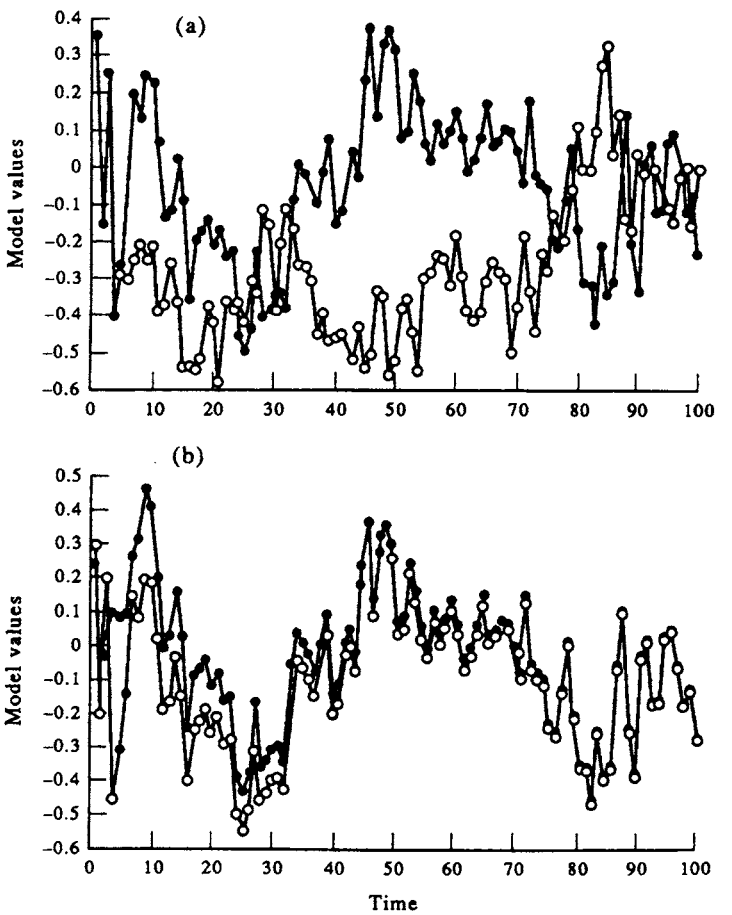


Figure 5.7.5. Two simulated realizations from the AR(4) model given by equation (5.7.25). (a) Same starting values but different and independently derived noise sequence; (b) different starting values but the same noise sequence. (From Gray and Woodward, 1992.)

procedure (Figure 5.7.6). As the figure illustrates, the maximum entropy spectrum has narrow peaks while both the Bartlett window and maximum likelihood method yield much broader spectral peaks. Note also that, except for the maximum spectral values, the maximum entropy spectrum significantly underestimates the spectral estimates for the 0.15 Hz signal and white noise. The maximum entropy spectrum also has small side-lobe energy that is dramatically less than the off-peak energy in either of these two spectra. The maximum likelihood spectral values are also systematically lower than those using the standard method with a Bartlett window. A similar comparison is shown in Figure 5.7.7, which first shows a time series of a 1 Hz (1 cps) sinusoid with 10% white noise added to it (Figure 5.7.7a). The power spectrum computed as the square of the Fourier coefficients is displayed in Figure 5.7.7(b). This can be compared with the narrow-peaked maximum entropy spectrum in Figure 5.7.7(c). The peaks are located at the same frequency representative of the 1 Hz, but the maximum entropy spectrum is extremely narrow while the Fourier power spectrum has a very wide peak. It is easy to see that the maximum entropy method seriously underestimates the spectral values at frequencies other than the main peak.

5.8 CROSS-SPECTRAL ANALYSIS

Estimation of autospectral density functions deals only with the frequency characteristics of a single scalar or vector time series, $x(t)$. Estimation of cross-spectral density functions performs a similar analysis but for two time series, $x_1(t)$ and $x_2(t)$, spanning concurrent times, $0 \leq t \leq T$. Although we often use time series from similar distri-

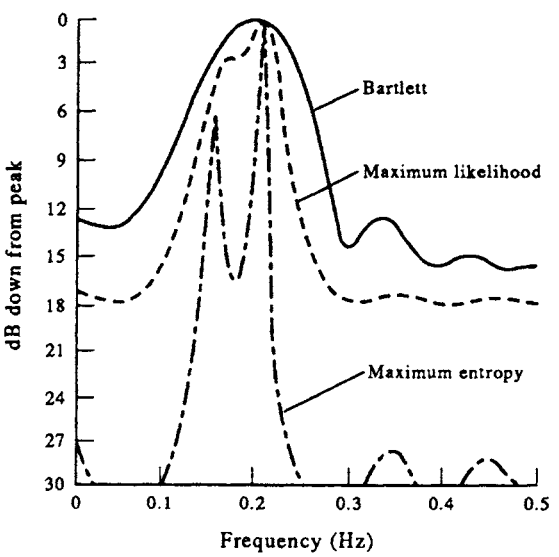


Figure 5.7.6. Power spectral estimates for a signal consisting of white noise plus two sine waves with frequencies 0.15 and 0.2 Hz (cps). Solid line: spectrum using the autocovariance method with a Bartlett smoothing window. Dashed line: Maximum likelihood spectral estimate. Dash–dot line: maximum entropy spectrum. (From Lacoss, 1971.)

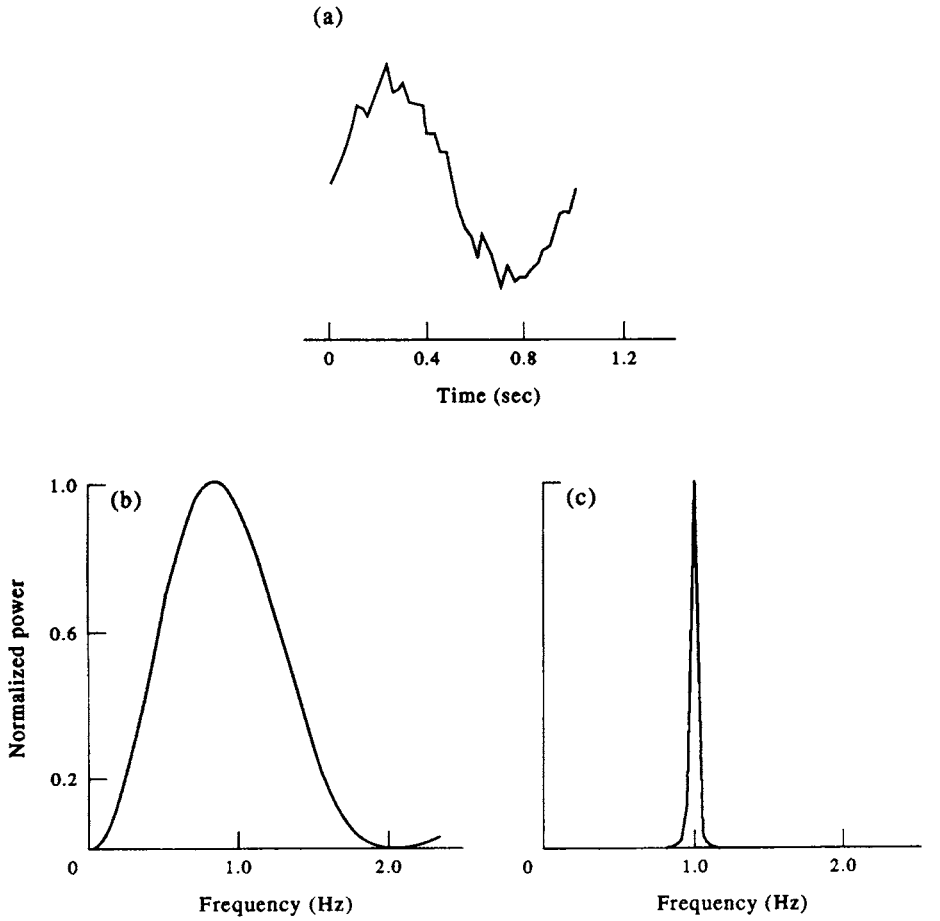


Figure 5.7.7. Comparison of spectra from periodogram method and maximum entropy method (MEM). (a) A sinusoid with 10% white noise and truncated with a 1 s window; (b) the power spectrum of (a) computed as the square of the modulus of the Fourier transform; (c) the MEM power spectrum of (a). Frequency in Hz (cps). (From Ulrych, 1972.)

butions, such as the velocity records from nearby moorings, cross-spectra may also be computed for two completely different quantities. In that sense, we *can* mix apples and oranges. For example, the cross-spectrum formed from the time-varying velocity fluctuations, $x_1(t) = u'(t)$, and the temperature fluctuations, $x_2(t) = T'(t)$, measured over the same time span at the same location gives an estimate of the local eddy heat flux, $q' = \rho C_\rho u'T'(t)$, as a function of frequency (ρ is the density and C_ρ the specific heat of seawater). Because autospectra involve terms like $x_1 x_1^*$, where the asterisk denotes complex conjugate, the spectra are real-valued and all phase information in the original signal is lost. Cross-spectra, on the other hand, involve terms like $x_1 x_2^*$ and are generally complex quantities whose real and imaginary parts take into account the correlated portions of both the amplitudes and relative phases of the two signals.

There are two ways to quantify the real and imaginary parts of cross-spectra. One approach is to write the cross-spectrum as the product of an amplitude function, called the *cross-amplitude spectrum*, and a phase function called the *phase spectrum*. The sample cross-amplitude spectrum gives the distribution of co-amplitudes with frequency while the sample phase spectrum indicates the angle (or time) by which one

series leads or lags the other series as a function of frequency. Alternatively, the cross-spectrum can be decomposed into a *coincident spectral density function (or co-spectrum)*, which defines the degree of co-oscillation for those frequency constituents of the two time series that fluctuate in-phase, and a *quadrature spectral density function (or quad-spectrum)*, which defines the degree of co-oscillation for frequency constituents of the two series that co-oscillate but are out-of-phase by $\pm 90°$. Statistical confidence intervals can be provided for normalized versions of the cross-spectral estimates.

5.8.1 Cross-correlation functions

In Section 5.6.3.1, we showed that the autocovariance function, $C_{xx}(\tau)$, and the autospectrum, $S_{xx}(f)$, are Fourier transform pairs. Similarly, for separate time series $x_1(t)$ and $x_2(t)$, the cross-covariance function, $C_{x1x2}(\tau)$, and the cross-spectrum, $S_{x1x2}(f)$, are transform pairs. Thus, we can take the Fourier transform of the lagged cross-covariance function to obtain the cross-spectrum or we can take the inverse Fourier transform of the cross-spectrum to obtain the cross-covariance function. As a prelude to cross-spectral analysis, it is worth presenting a brief summary of cross-correlation functions commonly used in oceanography for scalar and vector time series. The cross-correlation functions tell us how closely two records are "related" in the time domain, whereas the cross-spectrum tells us how oscillations within specific frequency bands are related in the frequency domain.

Using the abbreviation $C_{12}(\tau)$ for $C_{x1x2}(\tau)$, the *cross-covariance function* is defined as

$$C_{12}(\tau) = \frac{1}{N-m}\sum_{N=0}^{N-m} x_1(n\Delta t)x_2(n\Delta t + \tau) \tag{5.8.1}$$

where $\tau = m\Delta t$ is the lag time for $m = 0, 1, \ldots, M$, $M \ll N$. Division of (5.8.1) by the product $C_{11}(0)C_{22}(0)$, corresponding to the autocovariance functions for each series at zero lag, gives the *cross-correlation coefficient function* for the data samples

$$\rho_{12}(\tau) = \frac{C_{12}(\tau)}{[C_{11}(0)C_{22}(0)]^{1/2}} \tag{5.8.2}$$

The time series $x_1(t)$ and $x_2(t)$ represent any two quantities we wish to compare. They also may represent quantities measured at different depths or locations for the same time period. For example, Kundu and Allen (1976) used the lagged covariance function

$$\begin{aligned}\rho(\mathbf{x}_1, \mathbf{x}_2, \tau) &= \frac{\overline{v'(\mathbf{x}_1, t)v'(\mathbf{x}_2, t+\tau)}}{\left[(v'(\mathbf{x}_1, t))^2 (v'(\mathbf{x}_2, t))^2\right]^{1/2}} \\ &= \frac{\dfrac{1}{N-m}\displaystyle\sum_{n=1}^{N-m} v'(\mathbf{x}_1, n)v'(\mathbf{x}_2, n+m)}{\dfrac{1}{N}\left[\displaystyle\sum_{n=1}^{N}\Big(v'(\mathbf{x}_1, n)\Big)^2\Big(v'(\mathbf{x}_2, n)\Big)^2\right]^{1/2}}, \quad m = 0, 1, \ldots, M \ll N\end{aligned} \tag{5.8.3}$$

to examine the correlation between the longshore (v) components of current for different coastal sites separated by a distance $d = |\mathbf{x}_1 - \mathbf{x}_2|$. Moreover, if $\tau_{\max}$ is the

lag which gives the maximum correlation, then the speed of propagation, c, of the coherent signal in the direction $\mathbf{d} = \mathbf{x}_1 - \mathbf{x}_2$ is $c = |\mathbf{d}|/\tau_{\max}$, the direction of propagation determined from the sign of $\tau_{\max}$ (Figure 5.8.1). In Figure 5.8.1, the lagged correlations between time series of low-pass filtered longshore currents, $v(\mathbf{x}, t)$, at different sites along the continental shelf are used to examine the poleward propagation of low-frequency coastal-trapped waves. Results in the figure are based on currents at 60-m depth. Letters refer to pairs of stations used; e.g. C – P is the lag between the Carnation and Poinsettia stations.

A generalization of (5.8.3) is given by Kundu (1976). If $w = u + iv$ is the complex velocity, then the correlation between the rotating velocity vectors is given by the complex correlation coefficient

$$\rho(\mathbf{x}_1, \mathbf{x}_2, \tau) = \frac{\overline{w_1^*(t) w_2(t+\tau)}}{\left[\overline{w_1^*(t) w_1(t)}^{1/2} \overline{w_2^*(t) w_2(t)}^{1/2}\right]} \tag{5.8.4}$$

where subscripts denote locations 1 and 2, and the overbars denote the time or ensemble average. The correlation, ρ, which is independent of the choice of coordinate systems, is a complex quantity whose magnitude gives the overall measure of correlation and whose phase gives the average counterclockwise angle of the second vector with respect to the first.

5.8.2 Cross-covariance method

Following the Blackman–Tukey procedure for autospectral density estimation, the Fourier transform of the cross-covariance function, $C_{12}(\tau)$, can be used to find the cross-spectrum, $S_{12}(f)$. Although the cross-covariance method is straightforward to apply, the sample cross-covariance function, $C_{12}(\tau)$, suffers from the same disadvantage as the sample autocovariance function, $C_{11}(\tau)$, in that neighboring values tend to be highly correlated, thereby reducing the effective number of degrees of freedom. Moreover, the statistical significance falls off rapidly with increasing lag, τ, so that the number of lags, M, is much shorter than the record length ($M \ll N$).

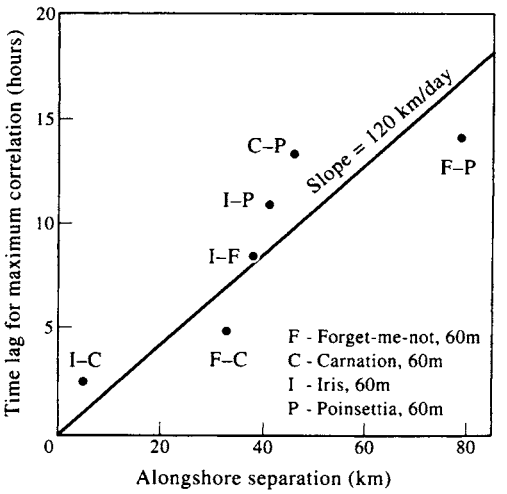


Figure 5.8.1. The lag time of maximum correlation of the longshore component of current at 60-m depth versus the distance of separation for the Oregon coast for 1973. Results indicate a mean northward signal propagation of 120 km/day. (From Kundu and Allen, 1976.)

Calculation of cross-spectra is best performed using the direct Fourier transform method. In fact, it is common practice these days to use the inverse Fourier transform of the cross-spectrum to get the cross-covariance function.

5.8.3 Fourier transform method

As with autospectral analysis, estimates of cross-spectral density functions are most commonly derived using Fourier transforms. The steps in calculating the cross-spectrum using standard Fourier transforms or FFTs are as follows (see also Bendat and Piersol, 1986):

(1) Ensure that the two time series $x_1(t)$ and $x_2(t)$ span the same period of time, t_n, where $n = 0, 1, \ldots, N-1$, and $T = N\Delta t$ is the length of each record. Remove the means and trends from each of the two time series. If block averaging is to be used to improve the statistical reliability of the spectral estimates, divide the available data for each pair of time series into m sequential blocks of N' data values each, where $N' = N/m$

(2) To reduce side-lobe leakage, taper the time series $x_1(t)$ and $x_2(t)$ using a Hanning (raised-cosine) window, Kaiser–Bessel window, or other appropriate taper. Re-scale the spectra calculated in step 4 to account for the loss of "energy" during application of the window (see Table 5.6.4).

(3) Compute the Fourier transforms, $X_1(f_k)$, $X_2(f_k)$, $k = 0, 1, 2, \ldots, N-1$, for the two time series $x_1(t)$ and $x_2(t)$. For block-segmented data, calculate the Fourier transforms $X_{1m}(f_k)$ and $X_{2m}(f_k)$ for each of the m blocks, where $k = 0, 1, \ldots, N'-1$. To reduce the variance associated with the tapering in step 2, the transforms can be computed for overlapping segments.

(4) Adjust the scale factor of $X_1(f_k)$ and $X_2(f_k)$ [or $X_{1m}(f_k)$, $X_{2m}(f_k)$] for the reduction in spectral energy due to the tapering in step 2. For the Hanning window, multiply the amplitudes of the Fourier transforms by $\sqrt{(8/3)}$.

(5) Compute the raw cross-spectral power density estimates for each pair of time series (or each pair of blocks) where for the two-sided spectral density estimate

$$S_{12}(f_k) = \frac{1}{N\Delta t}[X_1^*(f_k)X_2(f_k)], \quad k = 0, 1, 2, \ldots, N-1$$

(no block averaging)

$$S_{12}(f_k;\ m) = \frac{1}{N\Delta t}[X_{1m}^*(f_k)X_{2m}(f_k)], \quad k = 0, 1, 2, \ldots, N'-1 \tag{5.8.5a}$$

(to be used for block averaging)

and for the one-sided spectral density estimates

$$G_{12}(f_k) = \frac{2}{N\Delta t}[X_1^*(f_k)X_2(f_k)], \quad k = 0, 1, 2, \ldots, N/2$$

(no block averaging)

$$G_{12}(f_k; m) = \frac{2}{N\Delta t}[X_{1m}^*(f_k)X_{2m}(f_k)], \quad k = 0, 1, 2, \ldots, N'/2 \tag{5.8.5b}$$

(for block averaging)

(6) In the case of the block-segmented data, average the raw cross-spectral density estimates from the m blocks of data to obtain the smoothed periodogram for $S_{12}(f_k)$, the two-sided cross-spectrum, or $G_{12}(f_k)$, the one-sided cross-spectrum.

Cross-covariance function: Since the cross-covariance function, $C_{12}(\tau)$ $[= R_{12}(\tau)$, the cross-correlation function, if the mean is removed from the record], and the cross-spectrum are Fourier transform pairs, equation (5.8.5) can be used to obtain a smoothed or unsmoothed estimate of the cross-covariance function. To do this, we first calculate the Fourier transforms $X_1(f)$ and $X_2(f)$ of the individual time series, and determine the product $S_{12}(f) = (N\Delta t)^{-1}[X_1^*(f)X_2(f)]$. We then take the inverse Fourier transform (IFT) of the cross-spectrum, $S_{12}(f)$, to obtain the cross-covariance function

$$C_{12}(\tau) = \int_{-\infty}^{\infty} S_{12}(f)e^{i2\pi f\tau}\, df \tag{5.8.6}$$

If the spectrum is unsmoothed prior to the IFT (or IFFT if the number of spectral estimates is a power of 2), we obtain the raw cross-covariance function. If, on the other hand, the cross-spectrum is smoothed prior to (5.8.6) using one of the spectral windows, such as the Hanning window, the cross-covariance function also will be a smoothed function.

We can use the acoustic backscatter data in Table 5.1(a) to illustrate the direct and indirect methods for calculating the cross-covariance function. In Table 5.8.1, we present the normalized, unsmoothed cross-covariance function, $\rho_{12}(\tau) = C_{12}(\tau)/[C_{11}(0)C_{22}(0)]^{1/2}$, obtained directly from the definition (5.8.1). In this case, the lag τ is in 5-m depth increments. The indirect approach is based on the Fourier estimates presented in Table 5.8.2. Here, we first give the Fourier transforms, $X_1(f)$ and $X_2(f)$, of the two profile series as a function of wavenumber, f (Table 5.8.2a). We next calculate the cross-spectrum, $S_{12}(f) = (N\Delta t)^{-1}[X_1^*(f)X_2(f)]$, and then take the inverse transform of $S_{12}(f)$ to obtain the cross-covariance function $C_{12}(\tau)$ as a function of lag (Table 5.8.2b). No smoothing was applied to either data set, and the results obtained from the inverse Fourier transform method are identical to those listed in Table 5.8.1, within roundoff error. The advantage of the transform approach is that it is straightforward to derive a smoothed cross-covariance function by windowing the cross-spectral estimate prior to Fourier inversion.

5.8.4 Phase and cross-amplitude functions

Suppose that the constituents of the bivariate time series $\{x_1(t), x_2(t)\}$ have the same frequency, f_0, but different amplitudes (A_1, A_2) and different phases (ϕ_1, ϕ_2), respectively. In particular, let

$$x_k(t) = A_k \cos(2\pi f_0 t + \phi_k), \quad k = 1, 2 \tag{5.8.7}$$

The Fourier transform of $x_k(t)$, over $-T/2 \le t \le T/2$ is

$$X_k(f) = \frac{A_k}{2}\left\{e^{i\phi_k}\frac{\{\sin[\pi(f-f_0)T]\}}{\pi(f-f_0)} + e^{-i\phi_k}\frac{\{\sin[\pi(f+f_0)T]\}}{\pi(f+f_0)}\right\}, \quad i = 1, 2 \tag{5.8.8}$$

Hence, the sample cross-spectra of the two series is

$$S_{12}(f) = \frac{1}{T}[X_1^*(f)X_2(f)] \tag{5.8.9}$$

where X_1^* is the complex conjugate of X_1. From this expression, we obtain

$$\begin{aligned} S_{12}(f) = \frac{A_1A_2}{4T}\left\{e^{-i\phi_1}\frac{\sin[\pi(f-f_0)T]}{\pi(f-f_0)} + e^{i\phi_1}\frac{\sin[\pi(f+f_0)T]}{\pi(f+f_0)}\right\} \\ \times\left\{e^{i\phi_2}\frac{\sin[\pi(f-f_0)T]}{\pi(f-f_0)} + e^{-i\phi_2}\frac{\sin[\pi(f+f_0)T]}{\pi(f+f_0)}\right\} \end{aligned} \tag{5.8.10}$$

where

Table 5.8.1. Unsmoothed, normalized cross-covariance function, $\rho_{12}(\tau)$ given by (5.8.2), as a function of lag τ in increments of 5 m for bin 1 of beams 1 and 2 of the acoustic backscatter spatial series (profiles) listed in Table 5.1(a)

Lag τ (m)	0	5	10	15	20	25	30	35
	0.96	0.94	0.85	0.71	0.57	0.48	0.40	0.31
Lag τ (m)	40	45	50	55	60	65	70	75
	0.23	0.14	0.02	−0.19	−0.24	−0.37	−0.46	−0.48

Table 5.8.2(a) Complex Fourier transforms of $X_1(f_k)$ and $X_2(f_k)$ for the profiles of acoustic backscatter listed in Table 5.1(a). For each wavenumber, f_k, the table lists the real part of the transform (top) followed by the imaginary part (bottom), where $X_j(f_k) = ReX_j(f_k) + iImX_j(f_k)$, $j = 1, 2$. The vertical wavenumber $f_k = kf'$, $k = 0, 1, \ldots, 16$, where the fundamental vertical wavenumber, $f' = 1/155$ m $= 0.00645$ cpm (cycles per meter)

FFT $k = 0$	1	2	3	4	5	6	7
$X_1(f_k)$ 348.13	−289.32	71.17	15.52	55.16	97.59	−28.66	5.07
0.00	214.96	−16.35	−117.25	105.57	−16.98	−21.37	4.28
$X_2(f_k)$ 339.02	−226.53	119.54	55.84	−5.24	59.55	−36.39	4.22
0.00	227.88	38.22	−93.12	122.33	−24.13	−6.57	−19.09

$k = 8$	9	10	11	12	13	14	15	16
1.13	−6.16	41.11	24.03	−1.79	4.63	3.74	4.09	27.13
6.87	21.29	−2.96	−36.43	−4.60	1.08	3.54	18.45	0.00
11.90	5.68	23.89	13.85	3.96	7.37	11.27	2.34	27.79
−5.35	−4.63	−5.13	−18.72	−1.67	−4.93	−4.47	9.00	0.00

Table 5.8.2(b) The inverse fast Fourier transform (IFFT) of the cross-spectrum $S_{12}(f_k) = (N\Delta t)^{-1}[X_1^*(f_k)X_2(f_k)]$ using the values in Table 5.8.2(a). The values represent the raw (unnormalized) estimates of the cross-covariance function, $C_{12}(\tau)$, as a function of lag $\tau(0 \le \tau \le 16)$ in increments of 5 m for bin 1 of beams 1 and 2 of the acoustic backscatter spatial series (profiles) listed in Table 5.1(a)

	$\tau = 0$	1	2	3	4	5	6	7
$C_{12}(\tau)$	13483.7	12752.4	11151.5	9087.4	6992.3	5436.5	4589.9	3411.7

$\tau = 8$	9	10	11	12	13	14	15	16
2382.5	1393.8	160.6	−1103.6	−2096.0	−3103.5	−3610.5	−3623.8	−3222.1

$$S_{12}(f) \underset{T\to\infty}{\to} \frac{A_1A_2}{4}\left[e^{-i(\phi_2-\phi_1)}\delta(f+f_0) + e^{i(\phi_2-\phi_1)}\delta(f-f_0)\right] \tag{5.8.11}$$

The phase difference, $(\phi_2 - \phi_1)$, in the above expressions determines the lead (or lag) of one cosine oscillation relative to the other for given frequency, f. The cross amplitude, A_1A_2, gives the geometric mean amplitude of the co-oscillation for frequency f. From equation (5.8.2), the sample cross-spectrum is

$$S_{12}(f) = \frac{A_1(f)A_2(f)}{T}\left[e^{i[\phi_2(f)-\phi_1(f)]}\right] \tag{5.8.12}$$

or

$$S_{12}(f) = A_{12}(f)\left[e^{i\phi_{12}(f)}\right] \tag{5.8.13}$$

where the sample phase spectrum, $\phi_{12}(f) = \phi_2(f) - \phi_1(f)$, is an odd function of frequency, and the sample cross-amplitude spectrum, $A_{12}(f) = A_1(f)A_2(f)/T$, is a positive even function of f.

5.8.5 Coincident and quadrature spectra

An alternative description of this same information is to describe cross-spectra in terms of coincident (C) and quadrature (Q) spectra. In this case, we can write

$$S_{12}(f) = C_{12}(f) - iQ_{12}(f) \tag{5.8.14}$$

where

$$C_{12}(f) = A_{12}(f)\cos[\phi_{12}(f)];\ Q_{12}(f) = -A_{12}(f)\sin[\phi_{12}(f)] \tag{5.8.15}$$

and

$$A_{12}^2(f) = C_{12}^2(f) + Q_{12}^2(f);\ \phi_{12}(f) = \tan^{-1}\left[\frac{-Q_{12}(f)}{C_{12}(f)}\right] \tag{5.8.16}$$

Here $C_{12}(f)$ is an even function of frequency and $Q_{12}(f)$ is an odd function. (The co-spectral density function $C_{12}(f)$ for frequency f is not to be confused with the covariance function $C_{12}(\tau)$ at time lag τ. Where confusion may arise, we use the cross-correlation $R_{12}(\tau)$ in place of $C_{12}(\tau)$.) If we consider the bivariate cosine example that we used in (5.8.7), we have

$$\begin{aligned} C_{12}(f) &= \frac{A_1A_2}{4}\cos(\phi_2-\phi_1)[\delta(f+f_0)+\delta(f-f_0)] \\ &= \left\{\frac{A_1\cos\phi_1 A_2\cos\phi_2}{4} + \frac{A_1\sin\phi_1 A_2\sin\phi_2}{4}\right\}[\delta(f+f_0)+\delta(f-f_0)] \end{aligned} \tag{5.8.17}$$

The sample co-spectrum, $C_{12}(f)$, measures the covariance between the two cosine components and the two sine components. That is, the contributions to the cross-spectrum from those components of the two time series that are "in phase" (phase differences of 0 or 180°). The sample quadrature spectrum, $Q_{12}(f)$, determines the contributions from those components of the time series that are coherent but "out of phase" (phase difference ±90°).

5.8.5.1 Relationship of co- and quad-spectra to cross-covariance

The inverse transform of the cross-spectrum gives the cross-covariance (cross-correlation)

$$\begin{aligned} R_{12}(\tau) &= \int_{-\infty}^{\infty} [C_{12}(f) - iQ_{12}(f)]e^{i2\pi f\tau}\, df \\ &= \int_{-\infty}^{\infty} C_{12}(f)\cos(2\pi f\tau)\, df + \int_{-\infty}^{\infty} Q_{12}(f)\sin(2\pi f\tau)\, df \end{aligned} \qquad \varsigma(5.8.18)$$

Since $C_{12}(f)$ is an even function, $R_{12}(0) = \int_{-\infty}^{\infty} C_{12}(f)\, df$. If we define

$$\begin{aligned} C_{12}(f) &= \int_{-T}^{T} R_{12}^{+}(\tau)\cos(2\pi f\tau)\, d\tau \\ Q_{12}(f) &= \int_{-T}^{T} R_{12}^{-}(\tau)\sin(2\pi f\tau)\, d\tau \end{aligned} \qquad (5.8.19)$$

then

$$\begin{aligned} R_{12}^{+}(\tau) &= \tfrac{1}{2}[R_{12}(\tau) + R_{12}(-\tau)] \text{ (the even part)} \\ R_{12}^{-}(\tau) &= \tfrac{1}{2}[R_{12}(\tau) - R_{12}(-\tau)] \text{ (the odd part)} \end{aligned} \qquad (5.8.20)$$

5.8.6 Coherence spectrum (coherency)

The *squared coherency*, *coherence-squared function*, or *coherence spectrum* between two time series $x_1(t)$ and $x_2(t)$ is defined for frequencies f_k, $k = 0, 1, \ldots, N-1$, as

$$\begin{aligned} \gamma_{12}^2(f_k) &= \frac{|G_{12}(f_k)|^2}{G_{11}(f_k)G_{22}(f_k)} \\ &= \frac{|S_{12}(f_k)|^2}{S_{11}(f_k)S_{22}(f_k)} \\ &= \frac{[C_{12}^2(f_k) + Q_{12}^2(f_k)]}{S_{11}(f_k)S_{22}(f_k)} \end{aligned} \qquad (5.8.21)$$

where $G_{11}(f_k)$ is the one-sided spectrum (confined to $f_k \geq 0$), $S_{11}(f_k) = \frac{1}{2}G_{11}(f_k)$ is the two-sided spectrum defined for all frequencies and $G_{12}(f_k)$ is the one-sided cross-spectrum. Here

$$0 \leq |\gamma_{12}^2(f_k)| \leq 1 \qquad (5.8.22)$$

and

$$\gamma_{12}(f_k) = |\gamma_{12}^2(f_k)|^{1/2} e^{-i\phi_{12}f_k} \qquad (5.8.23)$$

where $|\gamma_{12}^2(f_k)|^{1/2}$ is the modulus of the coherence function and $\phi_{12}(f_k)$ the phase lag between the two signals at frequency f_k (Figure 5.8.2). In the literature, both the squared coherency, γ_{12}^2, and its square root are termed "the coherence" so that there is often a confusion in meaning (Julian, 1975). To avoid any ambiguity, it is best to use

squared-coherency when conducting coherence analyses once the sign of the coherence function is determined. This has the added advantage that squared coherency represents the fraction of the variance in x_1 ascribable to x_2 through a linear relationship between x_1 and x_2. Two signals of frequency f_k are considered highly coherent and in phase if $|\gamma_{12}^2(f_k)| \approx 1$ and $\phi_{12}(f_k) \approx 0$, respectively (Figure 5.8.2). The addition of random noise to the functions x_1 and x_2 of a linear system decreases the coherence-squared estimate and increases the noisiness of the phase associated with the system parameters. Estimation of $\gamma_{12}^2(f_k)$ is one of the most difficult problems in time-series analysis since it is so highly noise dependent. We also point out that phase estimates generally become unreliable where coherency amplitudes fall below the 90–95% confidence levels for a given frequency.

The real part of the coherence function, $\gamma_{12}(f_k)$, lies between -1 and $+1$ while the squared-coherency is between 0 and $+1$. If the noise spectrum, $S_{\varepsilon\varepsilon}(f_k)$, is equal to the output spectrum, then the coherence function is zero. This says that white noise is incoherent, as required. Also, when $S_{\varepsilon\varepsilon}(f_k) = 0$, we have $\gamma_{12}^2(f_k) = 1$; that is the coherence is perfect if there is no spectral noise in the input signal. It is important to note that, if no spectral smoothing is applied, we are assuming that we have no spectral noise. In this case, the coherency spectrum will be unity for all frequencies, which is clearly not physically realistic. Noise can be introduced to the system by

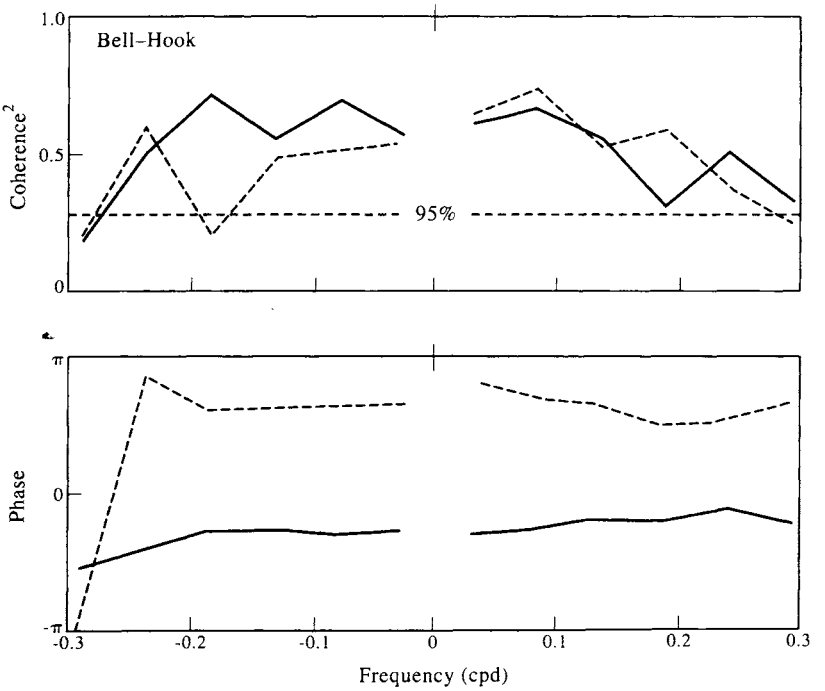


Figure 5.8.2. Coherence between current vector time series at sites Hook and Bell on the northeast coast of Australia (separation distance ≈ 300 km). (a) Coherence squared; (B) phase lag. Solid line: Inner rotary coherence (rotary current components rotating in the same sense). Dashed line: Outer rotary coherence (rotary current components rotating in the opposite sense). The increase in inner phase with frequency indicates equatorward phase propagation. Positive phase means that Hook leads Bell. (From Middleton and Cunningham, 1984.)

smoothing over adjacent frequencies. We also can overcome this problem by a prewhitening step that introduces some acceptable noise into the spectra.

5.8.6.1 Confidence levels

The final step in any coherence analysis is to specify the confidence limits for the coherence-square estimates. If $1 - \alpha$ is the $(1 - \alpha)100\%$ confidence interval we wish to specify for a particular coherence function, then, for all frequencies, the limiting value for the coherence-square (i.e. the level up to which coherence-square values can occur by chance) is given by

$$\begin{aligned} \gamma^2_{1-\alpha} &= 1 - \alpha^{[1/(\mathrm{EDOF}-1)]} \\ &= 1 - \alpha^{[2/(\mathrm{DOF}-2)]} \end{aligned} \tag{5.8.24}$$

where EDOF = DOF/2 (called the *equivalent* degrees of freedom) is the number of independent cross-spectral realizations in each frequency band (Thompson, 1979). The commonly used confidence intervals of 90, 95, and 99% correspond to $\alpha = 0.10$, 0.05, and 0.01, respectively. As an example, suppose that each of our coherence estimates is computed from an average over three adjacent cross-spectral Fourier components, then EDOF = 3 (DOF = 6). The 95% confidence level for the squared coherence would then be $\gamma^2_{95} = 1 - (0.05)^{0.5} = 0.78$. Alternatively, if the cross-spectrum and spectra were first smoothed using a Hamming window spanning the entire width of the data series, the equivalent degrees of freedom are EDOF = 2.5164 $\times$ 2 = 5.0328 (Table 5.6.4) and the 95% confidence interval $\gamma^2_{95} = 1 - (0.05)^{0.6595} = 0.86$. For EDOF = 2, $\gamma^2_{1-\alpha} = 1 - \alpha$ so that the confidence level is equal to itself.

A useful reference for coherence significance levels is Thompson (1979). In this paper, the author tests the reliability of significance levels $\gamma^2_{1-\alpha}$ estimated from (5.8.24) with the coherence-square values obtained through the summations

$$\gamma^2(f) = \frac{\left| \sum_{k=1}^{K} X_{1k}(f) X^*_{2k}(f) \right|^2}{\sum_{k=1}^{K} |X_{1k}(f)|^2 \sum_{k=1}^{K} |X_{2k}(f)|^2} \tag{5.8.25}$$

In this expression, X_{1k} and X_{2k} are the Fourier transforms of the respective random time series $x_{1k}(t)$ and $x_{2k}(t)$ generated by a Monte Carlo approach, and the asterisk denotes the complex conjugate. The upper limit K corresponds to the value of EDOF in (5.8.24a). Because $\gamma^2(f)$ is generated using random data, it should reflect the level of squared coherency that can occur by chance. For each value of K, $\gamma^2(f)$ was calculated 1000 times and the resultant values sorted as 90th, 95th, and 99th percentiles. The operation was repeated 10 times and the means and standard deviations calculated. This amounts to a total of 20,000 Fourier transforms for each K (=EDOF). There is excellent agreement between the significance level derived from (5.8.24) and the coherence-square values for a white-noise Monte Carlo process (Table 5.8.3), lending considerable credibility to the use of (5.8.24) for computing coherence significance levels. The comparisons in Table 5.8.3 are limited to the 90 and 95% confidence intervals for $4 \leq K \leq 30$. Thompson (1979) includes the 99% interval and a wider range of K (EDOF) values.

Confidence intervals for coherence amplitudes, as well as for coherence phase, admittance, and other signal properties (see next section), can be derived using the data itself (Bendat and Piersol, 1986). Let $\hat{\varphi}$ be an estimator for φ, a continuous, stationary random process, and define the standard error or random error of sample values as

$$\text{random error} = \sigma[\hat{\varphi}] = (E[\hat{\varphi}^2] - E^2[\hat{\varphi}])^{1/2} \tag{5.8.26a}$$

and the root mean square (RMS) error as

$$\text{RMS error} = (E[(\hat{\varphi} - \varphi)^2])^{1/2} = (\sigma^2[\hat{\varphi}] + B^2[\hat{\varphi}])^{1/2} \tag{5.8.26b}$$

where B is the bias term $B[\hat{\varphi}] = E[\hat{\varphi}] - \varphi$ and $E[x]$ is the expected value of x. If we now divide each error term by the quantity φ being estimated, we obtain the normalized random error

$$\varepsilon_r = \frac{\sigma[\hat{\varphi}]}{\varphi} = \frac{(E[\hat{\varphi}^2] - E^2[\hat{\varphi}])^{1/2}}{\varphi} \tag{5.8.27a}$$

and the normalized RMS error

$$\varepsilon = \frac{(E[(\hat{\varphi} - \varphi)^2])^{1/2}}{\varphi} = \frac{(\sigma^2[\hat{\varphi}] + B^2[\hat{\varphi}])^{1/2}}{\varphi} \tag{5.8.27b}$$

where it is assumed that $\varphi \neq 0$. Provided ε_r is small, the relation

$$\hat{\varphi}^2 = \varphi^2(1 \pm \varepsilon_r) \tag{5.8.28}$$

yields

$$\hat{\varphi} = \varphi(1 \pm \varepsilon_r)^{1/2} \approx \varphi(1 \pm \varepsilon_r/2) \tag{5.8.29}$$

so that

$$\varepsilon_r[\hat{\varphi}^2] \approx 2\varepsilon_r[\hat{\varphi}] \tag{5.8.30}$$

Thus, for small ε_r the normalized error for squared estimates $\hat{\varphi}^2$ is roughly twice the normalized error for unsquared estimates.

Table 5.8.3 Monte Carlo estimates, $\gamma^2(f)$, of the significant coherence-squared and prediction of this value using (5.8.24) for intervals $\alpha = 0.05$ and 0.10 for EDOF = 4, 5, 6, 8, 10, 20, and 30. (After Thompson, 1979)

	EDOF = 4	EDOF = 5	EDOF = 6	EDOF = 8	EDOF = 10	EDOF = 20	EDOF = 30
α =0.10							
$\gamma^2(f)$	0.539	0.437	0.371	0.288	0.230	0.114	0.076
$\gamma^2_{0.90}$	0.536	0.438	0.369	0.280	0.226	0.114	0.076
α =0.05							
$\gamma^2(f)$	0.629	0.531	0.452	0.354	0.288	0.144	0.099
$\gamma^2_{0.95}$	0.632	0.527	0.451	0.348	0.283	0.146	0.098

When the estimates $\hat{\phi}$ have a small bias error, $B[\hat{\varphi}] \approx 0$, and a small normalized error, e.g. $\varepsilon \leq 0.2$, the probability density for the estimates can be approximated by a Gaussian distribution. The confidence intervals for the unknown true parameter φ based on a single estimate $\hat{\varphi}$ are then

$$\hat{\varphi}(1-\varepsilon) \leq \varphi \leq \hat{\varphi}(1+\varepsilon) \text{ with 68\% confidence} \tag{5.8.31a}$$

$$\hat{\varphi}(1-2\varepsilon) \leq \varphi \leq \hat{\varphi}(1+2\varepsilon) \text{ with 95\% confidence} \tag{5.8.31b}$$

$$\hat{\varphi}(1-3\varepsilon) \leq \varphi \leq \hat{\varphi}(1+3\varepsilon) \text{ with 99\% confidence} \tag{5.8.31c}$$

5.8.7 Frequency response of a linear system

We define the admittance (or transfer) function of a linear system as

$$\begin{aligned} H_{12}(f_k) &= \frac{S_{12}(f_k)}{S_{11}(f_k)} = \frac{G_{12}(f_k)}{G_{11}(f_k)}, \quad f_k = k/T, \quad k = 1, \ldots, N \\ &= |H_{12}(f_k)| e^{-i\phi_{12}(f_k)} \end{aligned} \tag{5.8.32}$$

where $S_{11}(f_k)$ and $G_{11}(f_k)$ are, respectively, the two-sided and one-sided autospectrum estimates for the time series $x_1(t)$ selected here as the input time series. The gain (or admittance amplitude) function $H(f_k)$ behaves like a spectral regression coefficient at each frequency f_k. Using the definition $G_{12}(f_k) = C_{12}(f_k) - iQ_{12}(f_k)$, we obtain

$$\begin{aligned} |H_{12}(f_k)| &= \frac{G_{12}(f_k)}{G_{11}(f_k)} \\ &= \frac{|C_{12}^2(f_k) + Q_{12}^2(f_k)|^{1/2}}{G_{11}(f_k)} \end{aligned} \tag{5.8.33}$$

and where $\phi_{12}(f_k) = \tan^{-1}[-Q_{12}(f_k)/C_{12}(f_k)]$ by (5.8.16). Figure 5.8.3 shows the complex admittance for the observed longshore component of oceanic wind velocity (time series 1) and the longshore component of wind velocity derived from pressure-derived geostrophic winds (time series 2). The geostrophic winds closely approximate the amplitude and phase of the actual winds up to a frequency of about 0.05 cph (period = 20 h) after which the two signals no longer resemble one another. It is also at this frequency that the coherence consistently begins to fall below the 90% confidence level.

5.8.7.1 Multi-input systems cross-spectral analysis

Many oceanographic time series are generated through the combined effects of several mutually coherent inputs. For example, low-frequency fluctuations in coastal sea level typically arise through the combined forcing of atmospheric pressure, along- and cross-shore wind stress, and surface buoyancy flux. Coherences between the forcing variables (e.g. pressure, longshore wind stress, and runoff) are generally quite high. Because of this, it would be physically incorrect to use ordinary cross-spectral analysis which simply examines the correlation functions, $\gamma_{y:x}^2$, between the output, $y(t)$, and each of the inputs, $x(t)$, individually without taking into account the mutual correlation among all the inputs. If this is not done, the sum of the individual correlation functions can exceed unity. Provided that long-term sea-level fluctuations

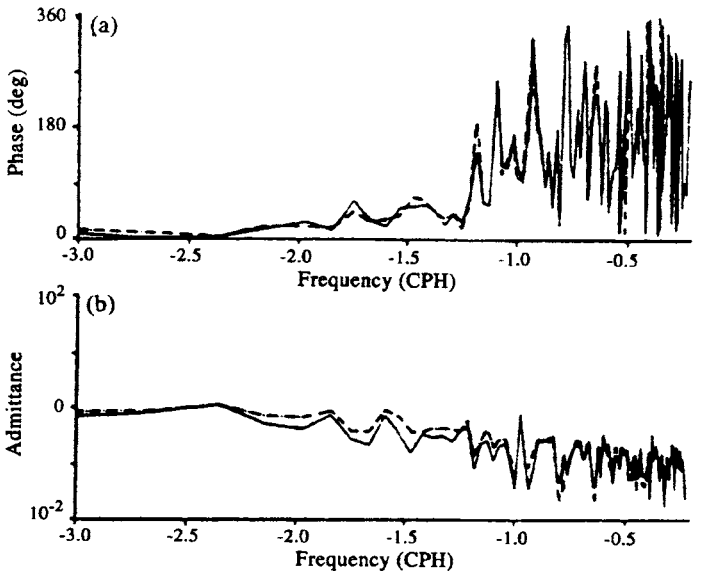


Figure 5.8.3. Complex admittance for observed (series 1) and calculated (series 2) longshore components of oceanic wind velocity. (a) Phase; (b) amplitude. Positive phase means that series 1 leads series 2. (From Thomson, 1983.)

(the output time series) are linearly related to the individual forcing functions (the input time series), we can use *multi-input systems cross-spectral analysis* to calculate the relative contribution each of the input terms makes to the output. The effective correlation function for the total system will then be less than unity, as required. This concept was pioneered in oceanography by Cartwright (1968), Groves and Hannan (1968), and Wunsch (1972). All three studies were concerned with sea-level variations.

The purpose of this section is to provide a brief overview of multiple systems analysis. For a thorough generalized presentation, the reader is directed to Bendat and Piersol (1986). Consider K constant-parameter linear systems associated with K stationary and ergodic input time series, $x_k(t)$, $k = 1, 2, \ldots, K$, a noise function, $\varepsilon(t)$, and a single output, $y(t)$, such that

$$y(t) = \sum_{k=1}^{K} y_k(t) + \varepsilon(t) \tag{5.8.34}$$

where the $y_k(t)$ are the outputs generated by each of the measured inputs $x_k(t)$. We can only measure the accumulated response $y(t)$, not the individual responses, $y_k(t)$. In the present context, $y(t)$ represents the measured time series of coastal sea level, $x_k(t)$ the corresponding weather variables, and $\varepsilon(t)$ the deviations from the ideal response due to instrument noise, remotely generated subinertial waves, and other physical processes not correlated with the input functions. The Fourier transform of the output $y(t)$ is

$$\begin{aligned} Y(f) &= \sum_{k=1}^{K} Y_k(f) + E(f) \\ &= \sum_{k=1}^{K} H_k(f) X_k(f) + E(f) \end{aligned} \tag{5.8.35}$$

where

$$H_k(f) = \frac{Y_k(f)}{X_k(f)}, \quad k = 1, 2, \ldots, K \tag{5.8.36}$$

is the admittance (or transfer) function relating the kth input with the kth output at frequency f. The frequency-domain spectral variables $X_k(f)$ and $Y(f)$ can be computed from the measured time series $x_k(t)$ and $y(t)$. Using these variables, we can then determine the functions $H_k(f)$ and other properties of the system.

Multiplication of both sides of (5.8.35) by $X_j^*(f)$, the complex conjugate of $X_j(f)$, for any fixed $j = 1, 2, \ldots, K$, yields the power spectral relation

$$S_{jy}(f) = \sum_{k=1}^{K} H_k(f) S_{jk}(f) + S_{j\varepsilon}(f), \quad j = 1, 2, \ldots, K \tag{5.8.37}$$

in which

$$\begin{aligned} S_{jy}(f) &= \overline{X_j^*(f) Y(f)}, \quad j = 1, 2, \ldots, K \\ S_{jk}(f) &= \overline{X_j^*(f) X_k(f)}, \quad j, k = 1, 2, \ldots, K \end{aligned} \tag{5.8.38}$$

Here, the overbar denotes the average value, the $S_{jy}(f)$ are the cross-spectra between the K inputs and the single output, $S_{jk}(f)$ are the cross-spectra ($j \neq k$) and spectra ($j = k$) among the input variables, and $S_{j\varepsilon}(f)$ is the cross-spectrum between the input variables and the noise function. If the noise function $\varepsilon(t)$ is uncorrelated with each input x_k (as is normally assumed), the cross-spectral terms $S_{j\varepsilon}(f)$ will be zero and (5.8.37) becomes

$$S_{jy}(f) = \sum_{k=1}^{K} H_k(f) S_{jk}(f), \quad j = 1, 2, \ldots, K \tag{5.8.39}$$

This expression is a set of K equations in K unknowns—the $H_k(f)$ for $k = 1, 2, \ldots, K$—where all spectral terms can be computed from the measured records of $y(t)$ and $x_k(t)$. If the model is well defined, matrix techniques can be used to find the $H_k(f)$. Bendat and Piersol (1986) also define the problem in terms of the *multiple and partial coherence functions* for the system. The multiple coherence function is given by

$$\gamma_{y:x}^2 = \frac{S_{vv}(f)}{S_{yy}(f)} = 1 - \frac{S_{\varepsilon\varepsilon}(f)}{S_{yy}(f)} \tag{5.8.40}$$

where $S_{vv}(f)$ is the multiple coherent output spectrum, $S_{yy}(f)$ is the output spectrum and $S_{\varepsilon\varepsilon}(f)$ is the noise spectrum. As with any squared coherence function, $0 \leq |\gamma_{y:x}^2| \leq 1$. For any problem with multiple inputs, $\gamma_{y:x}^2$ takes the form of a matrix whose off-diagonal elements take into account the coherent interactions among the different input terms. Expressions (5.8.39) and (5.8.40) simplify even further if the inputs themselves are mutually uncorrelated. In that case

$$H_j(f) = \frac{S_{jy}(f)}{S_{jj}(f)}, \quad j = 1, 2, \ldots, K; \quad |H_j(f)|^2 S_{jj}(f) = \gamma_{jy}^2 S_{yy}(f) \tag{5.8.41}$$

Hence, the contribution of the input variable, $x_j(t)$, to the output variable, $y(t)$, occurs

only through the transfer (admittance) function $H_j(f)$ of that particular input variable. No leakage of $x_j(t)$ takes place through any of the other transfer functions since $x_j(t)$ is uncorrelated with $x_k(t)$ for $k \neq j$.

In general, the output $y(t)$ is forced not only by the mutually coherent parts of the various inputs but also by the noncoherent portions of the inputs which go directly to the output through their own transfer functions without being affected by other transfer functions. This leads to the need for *partial coherence functions*. If part of one record causes part or all of a second record, then turning off the first record will eliminate the correlated parts from the second record and leave only that part of the second record that is not due to the first record. Because we do not want to incorporate the coherent portions of given forcing terms in the partial coherence functions, the partial coherences are found by first subtracting out the coherent parts of the various input signals. Bendat and Piersol (1986) state that, if any correlation between $x_1(t)$ and $x_2(t)$ is due to $x_1(t)$, then the optimum linear effects of $x_1(t)$ to $x_2(t)$ should be found. Denoting this mutual effect as $x_{2:1}(t)$, this should be subtracted from $x_2(t)$ to yield the conditioned (or residual) record, $x_{2:1}(t)$ representing that part of $x_2(t)$ not due to $x_1(t)$.

Multi-input systems cross-spectral analysis takes into account the fact that any input record $x_k(t)$ with nonzero correlations between other inputs will contribute to variations in the output $y(t)$ by passage through any of the K linear systems, $H_k(f)$. The conditioned portion of $x_k(t)$ will contribute directly to the output through its own response function only. The problem is to determine what percentage contribution each input function makes to the total variance of $y(t)$ for a specified frequency band. The simplest case is a two-input system consisting of inputs $x_1(t)$ and $x_2(t)$ for which

$$Y(f) = H_1(f)X_1(f) + H_2(f)X_2(f) + E(f) \tag{5.8.42}$$

and, provided $\gamma_{12}^2 \neq 0$

$$H_1(f) = \frac{S_{1y}(f)\left[1 - \dfrac{S_{12}(f)S_{2y}(f)}{S_{22}(f)S_{1y}(f)}\right]}{S_{11}(f)\left[1 - \gamma_{12}^2(f)\right]} \tag{5.8.43a}$$

$$H_2(f) = \frac{S_{2y}(f)\left[1 - \dfrac{S_{21}(f)S_{1y}(f)}{S_{11}(f)S_{2y}(f)}\right]}{S_{22}(f)\left[1 - \gamma_{12}^2(f)\right]} \tag{5.8.43b}$$

What is important to note here is the nonzero coupling between the different input variables when the cross-coherence, $\gamma_{12}^2(f)$, is nonzero. The product $H_1(f)S_{11}(f)$ in (5.8.43a) still represents the ordinary coherent spectrum between the input x_1 and the output y. However, when $|\gamma_{12}| \neq 0$, $x_1(t)$ influences $y(t)$ through the transfer function $H_2(f)$ as well as through its own transfer function $H_1(f)$. Similarly, $x_2(t)$ influences $y(t)$ through the transfer function $H_1(f)$ as well as through its transfer function $H_2(f)$ (5.8.43b). In general, the sum of $\gamma_{1y}^2(f)$ and $\gamma_{2y}^2(f)$ can be greater than unity when the outputs are correlated. The contributions from the conditioned records of $x_1(t)$ and $x_2(t)$ must also be taken into account when estimating the output response, $y(t)$. Once this is done, it becomes possible to construct reliable forecasting models for y.

Cartwright (1968) used the multiple input method to study tides and storm surges around east and north Britain. He expanded the tide height, ζ, at each of the ports studied as a Taylor series of the atmospheric pressure, p, about the port location ($x =$

$0, y = 0)$

$$\zeta(x,y,t) = p_{00}(t) + xp_{10}(t) + yp_{01}(t) + x^2p_{20}(t) + 2xyp_{11}(t) + y^2p_{02}(t) + \ldots \quad (5.8.44)$$

in which the pressure gradient terms $(p_{10}, p_{01}) = (\partial p/\partial x, \partial p/\partial y)$ are proportional to the geostrophic wind stress, the second derivatives $(p_{20}, p_{02}) = (\partial^2 p/\partial^2 x, \partial^2 p/\partial y^2)$ are related to wind stress gradients, and so on. As indicated by Table 5.8.4, the variances in different frequency bands for the sea level at Aberdeen, Scotland are significantly reduced relative to the original values as the pressure, first derivatives, and second derivatives are successively included. Consequently, all of the mutually correlated weather variables are considered relevant to the predictability of sea level. In a more recent study, Sokolova *et al.* (1992) used the multiple spectral analysis technique to study sea-level oscillations measured from July to September 1986 at different locations around the perimeter of the Sea of Japan. According to their analysis for both the multiple and partial coherences, 46–77% of the sea-level variance was coherent with atmospheric pressure and 5–37% was coherent with the wind stress.

5.8.8 Rotary cross-spectral analysis

As outlined in Section 5.6.4, the decomposition of a complex horizontal velocity vector, $w(t) = u(t) + iv(t)$, into counter-rotating circularly polarized components can aid in the analysis and interpretation of oceanographic time series. (Here, u and v typically represent the eastward and northward components of the current or wind.) Many of the fundamentals of this approach can be found in Fofonoff (1969), Gonella (1972), Mooers (1973), Calman (1978), and Hayashi (1979). In rotary spectral analysis, the different frequency components of the vector $w(t)$ are represented in terms of clockwise and counterclockwise rotating vectors (Figure 5.6.12). The counterclockwise component is considered to be rotating with positive angular frequency $(\omega \geq 0)$ and the clockwise component with negative angular frequency $(\omega \leq 0)$. Depending on which of the two components has the largest magnitude, the vector rotates clockwise or counterclockwise with time, with the tip of the vector tracing out an ellipse. If, for a given frequency, both components are of equal magnitude, the ellipse flattens to a line and the motions are *rectilinear* (back and forth along a straight line). Two one-sided autospectra and two one-sided cross-spectra can be computed for the rotary components. Mooers (1973) formulated these as two two-sided rotary autospectra called, respectively, the *inner* and *outer rotary autospectra*, the terminology originating from the resemblance of the inner and outer rotary autocovariance functions derived from the autospectra to the inner (dot) and outer (cross) products in mathematics. (A note

Table 5.8.4 Residual variances (cm^2) for different frequency bands for Aberdeen, Scotland sea-level oscillations. The predictive model explains increasingly more of the variance as additional weather variables are incorporated in the analysis. (Modified after Cartwright, 1968)

	0–0.5 cpd	0.5–0.8 cpd	1.1–1.8 cpd	2.1–2.8 cpd
Variables included				
Original variance	181	16	9.6	4.1
p_{00}	88	13	9.1	3.9
p_{00}, p_{10}, p_{01}	49	9	7.1	3.6
$p_{00}, p_{10}, p_{01}, \ldots p_{02}$	38	6	5.3	3.3

on terminology: Mooers (1973) uses A and C for counterclockwise (+) and clockwise components (−) while Gonella (1972) uses +/− subscripts for these components of the form u_+ and u_-. In this text, we use +/− superscripts where, for example, the amplitude of the two vector components is written as A^+ and A^-.)

To simplify the mathematics, we assume that u and v are continuous, stationary processes with zero means and Fourier integral representations. The velocity vector $w(t)$ can then be written in terms of its Fourier transform

$$w(t) = u(t) + iv(t) = \sum_p W_p e^{i\omega_p t}$$
$$= \sum_p \{[A_{1p}\cos(\omega_p t) + B_{1p}\sin(\omega_p t)] + i[A_{2p}\cos(\omega_p t) + B_{2p}\sin(\omega_p t)]\} \quad (5.8.45)$$

in which the Fourier transform component, W_p, is a complex quantity, the A and B are constants, and ω_p is the frequency of the pth Fourier component. As outlined in Section 5.6.4, each Fourier component of frequency $\omega = \omega_p$ can be expressed as a combination of two circularly polarized components having counterclockwise ($\omega \geq 0$) and clockwise ($\omega \leq 0$) rotation. Each of two components has its own amplitude and phase, and the tip of the vector formed by the combination of the two oppositely rotating components traces out an ellipse over a period, $T = 2\pi/\omega$. The semi-major axis of the ellipse has length $L_M = A^+(\omega) + A^-(\omega)$ and the semi-minor axis has length $L_m = |A^+(\omega) - A^-(\omega)|$. The angle, θ, of the major axis measured counterclockwise from the eastward direction gives the ellipse orientation.

If we specify $A_1(\omega)$ and $B_1(\omega)$ to be the amplitudes of the cosine and sine terms for the eastward (u) component in equation (5.8.45) and $A_2(\omega)$ and $B_2(\omega)$ to be the corresponding amplitudes for the northward (v) component, the amplitudes of the two counter-rotating vectors for a given frequency are

$$A^+(\omega) = \tfrac{1}{2}\left\{[B_2(\omega) + A_1(\omega)]^2 + [A_2(\omega) - B_1(\omega)]^2\right\}^{1/2} \quad (5.8.46a)$$

$$A^-(\omega) = \tfrac{1}{2}\left\{[B_2(\omega) - A_1(\omega)]^2 + [A_2(\omega) + B_1(\omega)]^2\right\}^{1/2} \quad (5.8.46b)$$

and their phases are

$$\tan(\theta^+) = [A_1(\omega) - B_1(\omega)]/[A_1(\omega) + B_2(\omega)] \quad (5.8.47a)$$

$$\tan(\theta^-) = [B_1(\omega) + A_2(\omega)]/[B_2(\omega) - A_1(\omega)] \quad (5.8.47b)$$

The eccentricity of the ellipse is

$$\varepsilon(\omega) = 2[A^+(\omega)A^-(\omega)]^{1/2}/[A^+(\omega) + A^-(\omega)] \quad (5.8.48)$$

where the ellipse traces out an area $\pi[(A^+)^2 - (A^-)^2]$ during one complete cycle of duration $2\pi/\omega$. The use of rotary components leads to two-sided spectra; i.e. defined for both negative and positive frequencies. If $S^+(\omega)$ and $S^-(\omega)$ are the rotary spectra for the two components, then $A^{\pm}(\omega) \propto [S^{\pm}(\omega)]^{1/2}$ can be used to determine the ellipse eccentricity. The sense of rotation of the vector about the ellipse is given by the rotary

coefficient (see Section 5.6.4.2)

$$r(\omega) = [S^+(\omega) - S^-(\omega)]/[S^+(\omega) + S^-(\omega)] \tag{5.8.49}$$

where $-1 \leq r \leq 1$. Values for which $r > 0$ indicate counterclockwise rotation while values of $r < 0$ indicate clockwise rotation; $r = 0$ is rectilinear motion.

If u, v are orthogonal Cartesian components of the velocity vector, $w = (u, v)$, then the rotary spectra can be expressed as

$$\begin{aligned} S^+(\omega) &= [A^+(\omega)]^2, \quad \omega \geq 0 \\ &= \tfrac{1}{2}[S_{uu} + S_{vv} + 2Q_{uv}] \end{aligned} \tag{5.8.50a}$$

$$\begin{aligned} S^-(\omega) &= [A^-(\omega)]^2, \quad \omega \leq 0 \\ &= \tfrac{1}{2}[S_{uu} + S_{vv} - 2Q_{uv}] \end{aligned} \tag{5.8.50b}$$

where S_{uu} and S_{vv} are the autospectra for the u and v components, and Q_{uv} is the quadrature spectrum between the two components. The stability of the ellipse is given by

$$\begin{aligned} \mu(\omega) &= \frac{|\langle (A^-(\omega)A^+(\omega) \exp[\mathrm{i}(\theta^+ - \theta^-)]\rangle|^2}{\langle (A^-)^2\rangle\langle (A^+)^2\rangle}, \quad \omega \geq 0 \\ &= \frac{|Y|}{[S^+(\omega)S^-(\omega)]^{1/2}} \end{aligned} \tag{5.8.51}$$

where

$$Y = \tfrac{1}{2}[S_{uu} - S_{vv} + \mathrm{i}2S_{uv}] \tag{5.8.52}$$

and the ellipse has a mean orientation

$$\phi = \tfrac{1}{2}\tan^{-1}[2S_{uv}/(S_{uu} - S_{vv})] \tag{5.8.53}$$

where ϕ is measured counterclockwise from east (the function ϕ is not coordinate invariant). The brackets $\langle \cdot \rangle$ denote an ensemble average or a band average in frequency space. The ellipse stability, $\mu(\omega)$, resembles the magnitude of a correlation function and is a measure of the confidence one might place in the estimate of the ellipse orientation (Gonella, 1972).

5.8.8.1 Rotary analysis for a pair of time series

Having summarized the rotary vector analysis for a single location, we now want to consider the coherence and cross-spectral properties for two time series measured simultaneously at two spatial locations. The object of the rotary spectral analysis is to determine the "similarity" between the two time series in terms of their circularly polarized rotary components. For two vector time series, the inner and outer rotary cross-spectra can be computed. As the spectra are complex, they have both amplitude and phase. Hence, coherence and phase spectra can be computed, just as with the cross-spectra of two scalar time series. *Inner* functions describe co-rotating compon-

ents and *outer* functions describe counter-rotating components. We could, of course, use standard Cartesian components for this task. Unfortunately, the Cartesian vectors and their derived relationships generally are dependent on the selected orientation of the coordinate system. The advantages of the rotary type of analysis are: (1) The coherence analysis is independent of the coordinate system (i.e. is coordinate invariant); and (2) the results encompass the coherence and phase of oppositely rotating, as well as like-rotating components, for motions that may be highly nonrectilinear. Because the counter-rotating components have circular symmetry, invariance under coordinate rotation follows for coherence.

We consider two vector time series defined by the relations

$$w_1(t) = (u_1, v_1); w_2(t) = (u_2, v_2) \tag{5.8.54}$$

where, as before, $(u, v) = u + iv$ are complex quantities. If $W_1(\omega)$ and $W_2(\omega)$ are components of the Fourier transforms of these time series, then the transforms can be expressed in the form

$$W(\omega) = \begin{cases} A^+ \exp(-i\theta^+), & \omega \geq 0 \\ A^- \exp(-i\theta^-), & \omega \leq 0 \end{cases} \tag{5.8.55}$$

with the same definitions for amplitudes and phases as in the previous subsection. These expressions equate the negative frequency components from the Fourier transform with the clockwise rotary components and the positive frequency components from the transform with the counterclockwise components.

Inner-cross spectrum: The inner cross-spectrum, $S_{wjwk}(\omega)$, provides an estimate of the joint energy content of two time series for rotary components rotating in the same direction (e.g. the clockwise component of series 1 with the clockwise component of series 2; Figure 5.8.4). For all frequencies, $-\omega_N < \omega < \omega_N$

$$\begin{aligned} S_{wjwk}(\omega) &= \left\langle W_j^*(\omega) W_k(\omega) \right\rangle, \quad j, k = 1, 2 \\ &= \begin{cases} A_j^+(\omega) A_k^+(\omega) \exp[-i(\theta_j^+ - \theta_k^+)], & \omega \geq 0 \\ A_j^-(\omega) A_k^-(\omega) \exp[i(\theta_j^- - \theta_k^-)], & \omega \leq 0 \end{cases} \end{aligned} \tag{5.8.56}$$

where, as before, $\langle \cdot \rangle$ denotes an ensemble average or a band average in frequency space, and the asterisk denotes the complex conjugate. It follows that the inner-autospectrum for each time series is

$$S_{w_j w_j}(\omega) = \begin{cases} [A_j^+(\omega)]^2, & \omega \geq 0 \\ [A_j^-(\omega)]^2, & \omega \leq 0 \end{cases} \tag{5.8.57}$$

Thus, $S_{w_j w_j}(\omega)$ $(j = 1, 2)$ is the power spectrum of the counterclockwise component of the series j for $\omega \geq 0$, and the power spectrum for the clockwise component for $\omega \leq 0$. The area under the curve of $S_{wjwk}(\omega)$ versus frequency equals the sum of the variance of the eastward (u) and northward (v) components. For $\omega \geq 0$, $S_{w_1 w_2}(\omega)$ is the cross-spectrum for the counterclockwise component of series 1 and 2, while for $\omega \leq 0$, $S_{w_1 w_2}(\omega)$ represents the cross-spectrum for the clockwise rotary component.

Inner-coherence squared: The two-sided inner-coherence squared, $\gamma_{12}^2(\omega)$, between the two time series at frequency ω is defined in the usual manner. Specifically, using the

previous definitions for the rotary components, we find

$$\gamma_{12}^2(\omega) = \begin{cases} \{\langle A_1^+ A_2^+ \cos(\theta_1^+ - \theta_2^+)\rangle^2 + \langle A_1^+ A_2^+ \sin(\theta_1^+ - \theta_2^+)\rangle^2\}/\langle A_1^{+2}\rangle\langle A_2^{+2}\rangle, & \omega \geq 0 \\ \{\langle A_1^- A_2^- \cos(\theta_1^- - \theta_2^-)\rangle^2 + \langle A_1^- A_2^- \sin(\theta_1^- - \theta_2^-)\rangle^2\}/\langle A_1^{-2}\rangle\langle A_2^{-2}\rangle, & \omega \leq 0 \end{cases} \tag{5.8.58}$$

where $0 \leq |\gamma_{12}^2| \leq 1$. A coherence of near zero indicates a negligible relationship between the two like-rotating series while a coherence near unity indicates a high degree of variability between the series. The inner-phase lag, ϕ_{12}, between the two vectors is

$$\phi_{12}(\omega) = \tan^{-1}[-\mathrm{Im}(S_{w_1 w_2})/\mathrm{Re}(S_{w_1 w_2})] \tag{5.8.59}$$

or, in terms of the clockwise and counterclockwise components

$$\tan(\phi_{12}) = \begin{cases} \langle A_1^+ A_2^+ \sin(\theta_1^+ - \theta_2^+)\rangle / \langle A_1^+ A_2^+ \cos(\theta_1^+ - \theta_2^+)\rangle, & \omega \geq 0 \\ \langle -A_1^- A_2^- \sin(\theta_1^- - \theta_2^-)\rangle / \langle A_1^- A_2^- \cos(\theta_1^- - \theta_2^-)\rangle, & \omega \leq 0 \end{cases} \tag{5.8.60}$$

The phase, which is the same for both the inner cross-spectrum and the inner coherence, is a measure of the phase lead of the rotary component of time series 1 with respect to that of time series 2. Figure 5.8.4(a) shows the inner rotary coherence and phase for five years of monthly winter (November through February) wind data measured off Alaska at Middleton Island (59.4°N, 146.3°W) and Environmental Weather Buoy EB03 (56.0°N, 148.0°W). Co-rotating wind vectors were generally

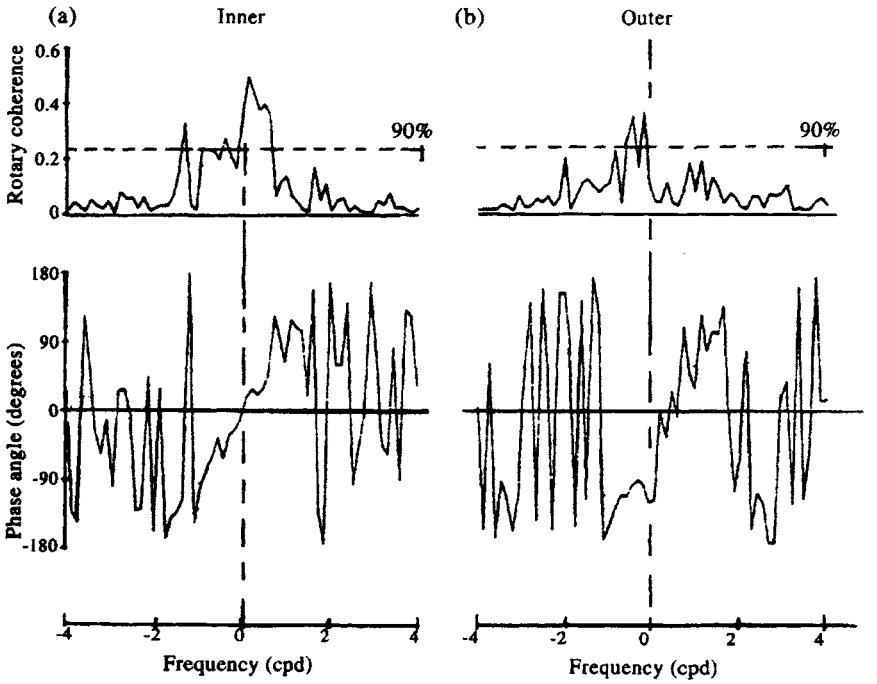


Figure 5.8.4. Rotary coherence and phase for five-year time series of monthly mean winter (November through February) wind velocity from two sites off Alaska. (a) Co-rotating (inner) coherence and phase with 90% confidence level; (b) counter-rotating (outer) coherence and phase. (From Livingstone and Royer, 1980.)

coherent above the 90% confidence level for frequencies $-1 < f < 1$ cpd, with greater coherence at positive frequencies (Livingstone and Royer, 1980). The inner phase was nearly a straight line in the frequency range $-1 < f < 0$ cpd, increasing by 120° over this range.

Outer-cross spectrum: The outer cross-spectrum, $Y_{w_j w_k}(\omega)$, provides an estimate of the joint energy content between rotary components rotating in opposite directions (e.g. between the clockwise component of time series 1 and the counterclockwise component of time series 2). For frequencies in the Nyquist frequency range, $-\omega_N < \omega < \omega_N$

$$\begin{aligned} Y_{wjwk}(\omega) &= \langle W_j(-\omega) W_k(\omega) \rangle, \quad j, k = 1, 2 \\ &= \begin{cases} A_j^-(\omega) A_k^+(\omega) \exp[\mathrm{i}(\theta_k^+ - \theta_j^-)], & \omega \geq 0 \\ A_j^+(\omega) A_k^-(\omega) \exp[\mathrm{i}(\theta_j^+ - \theta_k^+)], & \omega \leq 0 \end{cases} \end{aligned} \tag{5.8.61}$$

(Middleton, 1982). These relations resemble those for the inner-cross spectra but involve a combination of oppositely rotating vector amplitudes and phases. For the case of a single series, j, the outer rotary autospectrum is then

$$Y_{wjwj}(\omega) = A_j^-(\omega) A_j^+(\omega) \exp[\mathrm{i}(\theta_j^+ - \theta_j^-)], \quad \omega \geq 0 \tag{5.8.62}$$

and is symmetric about $\omega = 0$, and so is defined for only $\omega \geq 0$. Hence, $Y_{w_j w_j}(\omega)$ is an even function of frequency; i.e. $Y_{w_j w_j}(\omega) = Y_{w_j w_j}(-\omega)$. As noted by Mooers, $Y_{w_j w_j}(\omega)$ is not a power spectrum in the ordinary physical sense because it is complex valued. Rather it is related to the spectrum of the uv-Reynolds stress.

Outer-coherence squared: After first performing the ensemble or band averages in the brackets $\langle \cdot \rangle$, the outer-rotary coherence squared between series j and k is expressed in terms of the Fourier coefficients as

$$\lambda_{jk}^2(\omega) = \begin{cases} \langle A_j^- A_k^+ \rangle^2 \left[\langle \cos(\theta_k^+ - \theta_j^-) \rangle^2 + \langle \sin(\theta_k^+ - \theta_j^+) \rangle^2 \right] / \langle A_k^{+2} \rangle \langle A_j^{-2} \rangle, & \omega \geq 0 \\ \langle A_j^+ A_k^- \rangle^2 \left[\langle \cos(\theta_j^+ - \theta_k^-) \rangle^2 + \langle \sin(\theta_j^+ - \theta_k^+) \rangle^2 \right] / \langle A_j^{+2} \rangle \langle A_k^{-2} \rangle, & \omega \leq 0 \} \end{cases} \tag{5.8.63}$$

The phase lag, $\psi_{jk}(\omega)$ between the two oppositely rotating components of the two time series is then the same for the coherence and the cross-spectrum and is given by

$$\tan(\psi_{12}) = \begin{cases} \langle A_j^- A_k^+ \sin(\theta_j^- - \theta_k^+) \rangle / \langle A_j^- A_k^+ \cos(\theta_j^- - \theta_k^+) \rangle, & \omega \geq 0 \\ \langle A_j^+ A_k^- \sin(\theta_k^- - \theta_j^+) \rangle / \langle A_j^+ A_k^- \cos(\theta_k^- - \theta_j^+) \rangle, & \omega \leq 0 \end{cases} \tag{5.8.64}$$

If the values of

$$A_j^- A_k^+ \quad \text{and} \quad A_j^+ A_k^-$$

change little over the averaging interval covered by the angular brackets, then

$$\psi_{jk}(\omega) = \begin{cases} \theta_j^- - \theta_k^+, & \omega \geq 0 \\ \theta_k^- - \theta_j^+, & \omega \leq 0 \end{cases} \tag{5.8.65}$$

Figure 5.8.4(b) shows the outer rotary coherence and phase for five-year records of winter winds off Alaska. Counter-rotating vectors were coherent at negative frequencies in the range $-1 < f < 0$ cpd and exhibited little coherence at positive

frequencies. In this portion of the frequency band, the linear phase gradient was similar to that for the co-rotating vectors (Figure 5.8.4a).

Complex admittance function: If we think of the wind vector at location 1 as the source (or input) function and the current at location 2 as the response (or output) function, we can compute the complex inner admittance, Z_{12}, between two co-rotating vectors as

$$Z_{12}(\omega) = S_{w_1w_2}(\omega)/S_{w_1w_1}(\omega), \quad -\omega_N < \omega < \omega_N \tag{5.8.66}$$

The amplitude and phase of this function are

$$|Z_{12}(\omega)| = |S_{w_1w_2}(\omega)|/S_{w_1w_1}(\omega) \tag{5.8.67a}$$

$$\Phi_{12}(\omega) = \tan^{-1}\{\mathrm{Im}[S_{w_1w_2}(\omega)]/\mathrm{Re}[S_{w_1w_2}(\omega)]\} \tag{5.8.67b}$$

For frequency ω, the absolute value of $Z_{12}(\omega)$ determines the amplitude of the clockwise (counterclockwise) rotating response one can expect at location 2 to a given clockwise (counterclockwise) rotating input at location 1. The phase, $\Phi_{12}(\omega)$, determines the lag of the response vector to the input vector.

The corresponding expressions for the complex outer admittance, Z_{12}, between two opposite-rotating vectors are

$$Z_{12}(\omega) = Y_{w_1w_2}(\omega)/S_{w_1w_1}(\omega), \quad -\omega_N < \omega < \omega_N \tag{5.8.68}$$

with amplitude and phase

$$|Z_{12}(\omega)| = |Y_{w_1w_2}(\omega)|/S_{w_1w_1}(\omega) \tag{5.8.69a}$$

$$\Phi_{12}(\omega) = \tan^{-1}\{\mathrm{Im}[Y_{w_1w_2}(\omega)]/\mathrm{Re}[Y_{w_1w_2}(\omega)]\} \tag{5.8.69b}$$

For frequency ω, the absolute value of $Z_{12}(\omega)$ yields the amplitude of the clockwise (counterclockwise) rotating response one can expect at location 2 to a given counterclockwise (clockwise) rotating input at location 1. The phase, $\Phi_{12}(\omega)$, determines the lag of the response vector to the input vector.

5.9 WAVELET ANALYSIS

The terms "wavelet transform" and "wavelet analysis" are two recent additions to the lexicon of time-series analysis. First introduced in the 1980s for processing seismic data (cf. Goupillaud *et al.*, 1984), the technique has begun to attract attention in meteorology and oceanography where it has been applied to time-series measurements of turbulence (Farge, 1992; Shen and Mei, 1993), surface gravity waves (Shen *et al.*, 1994), low-level cold fronts (Gamage and Blumen, 1993), and equatorial Yanai waves (Meyers *et al.*, 1993).

As frequently noted in the literature, Fourier analysis does a poor job of dealing with signals of the form $\phi(t) = A(\tau)\cos(\omega t)$, where the amplitude, A, varies on the slow time scale, τ. Wavelet analysis has a number of advantages over Fourier analysis

that are particularly attractive. Unlike the Fourier transform, which generates record-averaged values of amplitude and phase for each frequency component or harmonic, ω, the wavelet transform yields a localized, "instantaneous" estimate for the amplitude and phase of each spectral component in the data set. This gives wavelet analysis an advantage in the analysis of nonstationary data series in which the amplitudes and phases of the harmonic constituents may be changing rapidly in time or space. Where a Fourier transform of the nonstationary time series would smear-out any detailed information on the changing processes, the wavelet analysis attempts to track the evolution of the signal characteristics through the data set. As with other transform techniques, problems can develop at the ends of the time series, and steps must be taken to mitigate these effects. Similar to other transform techniques involving finite length data, steps also must be taken to minimize the distortion of the transformed data caused by the nonperiodic behavior at the ends of the time series. Lastly, we note that increasing the temporal resolution, Δt, of the wavelet analysis decreases the frequency resolution, Δf, and vice versa, such that $\Delta t \Delta f < \frac{1}{4}\pi$, reminiscent of the Heisenberg uncertainty relation. The more accurately we want to resolve the frequency components of a time series, the less accurately we can resolve the changes in these frequency components with time.

5.9.1 The wavelet transform

Wavelet analysis involves the convolution of a real time-series, $x(t)$, with a set of functions $g_{a\tau}(t) = g(t{:}\,\tau, a)$ that are derived from a "mother wavelet" or analyzing wavelet, $g(t)$, which is generally complex. In particular

$$g_{a\tau}(t) = \frac{1}{\sqrt{a}} g[a^{-1}(t - \tau)] \tag{5.9.1}$$

where τ (real) is the *translation* parameter corresponding to the central point of the wavelet in the time series and a (real and positive) is the *scale dilation* parameter corresponding to the width of the wavelet. For the Gaussian-shaped Morlet wavelet (Figure 5.9.1) described in detail later in this section, the dilation parameter can be related to a corresponding Fourier frequency (or wavenumber).

The continuous wavelet transform, $X(t)$, of the time series with respect to the analyzing wavelet, $g(t)$, is defined through the convolution integral

$$X_g[\tau, a] = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} g^*[a^{-1}(t - \tau)] x(t)\, dt \tag{5.9.2}$$

in which g^* denotes the complex conjugate of g and variables τ, a are allowed to vary continuously through the domain $(-\infty, \infty)$. Wavelet analysis provides a two-dimensional unraveling of a one-dimensional time series into position, τ, and amplitude scale, a, as new independent variables. The wavelet transformation (5.9.2) is a sort of mathematical microscope, with magnification $1/a$, position τ, and optics given by the choice of the specific wavelet, $g(t)$ (Shen *et al.*, 1994). Whereas Fourier analysis provides an average amplitude over the entire time series, wavelet analysis yields a measure of the localized amplitudes a as the wavelet moves through the time series with increasing values of τ. Although wavelets have a definite scale, they typically do not bear any

resemblance to the sines and cosines of Fourier modes. Nevertheless, a correspondence between wavelength and scale a can sometimes be achieved.

To qualify for mother wavelet status, the function $g(t)$ must satisfy several properties (Meyers *et al.*, 1993):

(1) Its amplitude $|g(t)|$ must decay rapidly to zero in the limit $|t| \to \infty$. It is this feature that produces the localized aspect of wavelet analysis since the transformed values, $X_g[\tau, a]$ are generated only by the signal in the cone of influence about $t = \tau$. In most instances, the wavelet $g[(t - \tau)/a]$ is assumed to have an insignificant effect at some time $|t| = \tau_c$.

(2) $g(t)$ must have zero mean. Known as the *admissibility condition*, this ensures the invertability of the wavelet transform. The original signal can then be obtained from the wavelet coefficients through the inverse transform

$$x(t) = \frac{1}{C} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{X_g[\tau, a] a^{-2} g_{a\tau}\} \, d\tau \, da$$

where

$$C^{-1} = \int_{-\infty}^{\infty} (\omega^{-1} |G(\omega)|^2) \, d\omega \tag{5.9.4}$$

in which $G(\omega)$ is the Fourier transform of $g(t)$. For $1/C$ to remain finite, $G(0) = 0$.

(3) Wavelets are often regular functions, such that $G(\omega < 0) = 0$. These are also called *progressive* wavelets. Elimination of negative frequencies means that wavelets need only be described in terms of positive frequencies.

(4) Higher-order moments (such as variance and skewness) should vanish allowing the investigation of higher-order variations in the data. This requirement can be relaxed, depending on the application.

One of most extensively used wavelets is the standard (admissible and progressive) Morlet wavelet

$$g(t) = e^{-t^2/2} e^{+ict} \tag{5.9.5}$$

consisting of a plane wave of frequency $c = \omega$ (or wavenumber $c = k$ in the spatial domain) which is modulated by a Gaussian envelope of unit width. Another possible wavelet which is applicable to a signal with two frequencies c_1 and c_2 is

$$g(t) = e^{-t^2/2} e^{ic_1 t} e^{ic_2 t} \tag{5.9.6}$$

while the wavelet

$$g(t) = e^{-t^2/2} e^{ict} e^{ikt^2/2} \tag{5.9.7}$$

is applicable to short data segments with linearly increasing frequency ("chirps").

5.9.2 Wavelet algorithms

The choice of $g(t)$ is dictated by the analytical requirements. More specifically, the wavelet should have the same pattern or signal characteristic as the pattern being

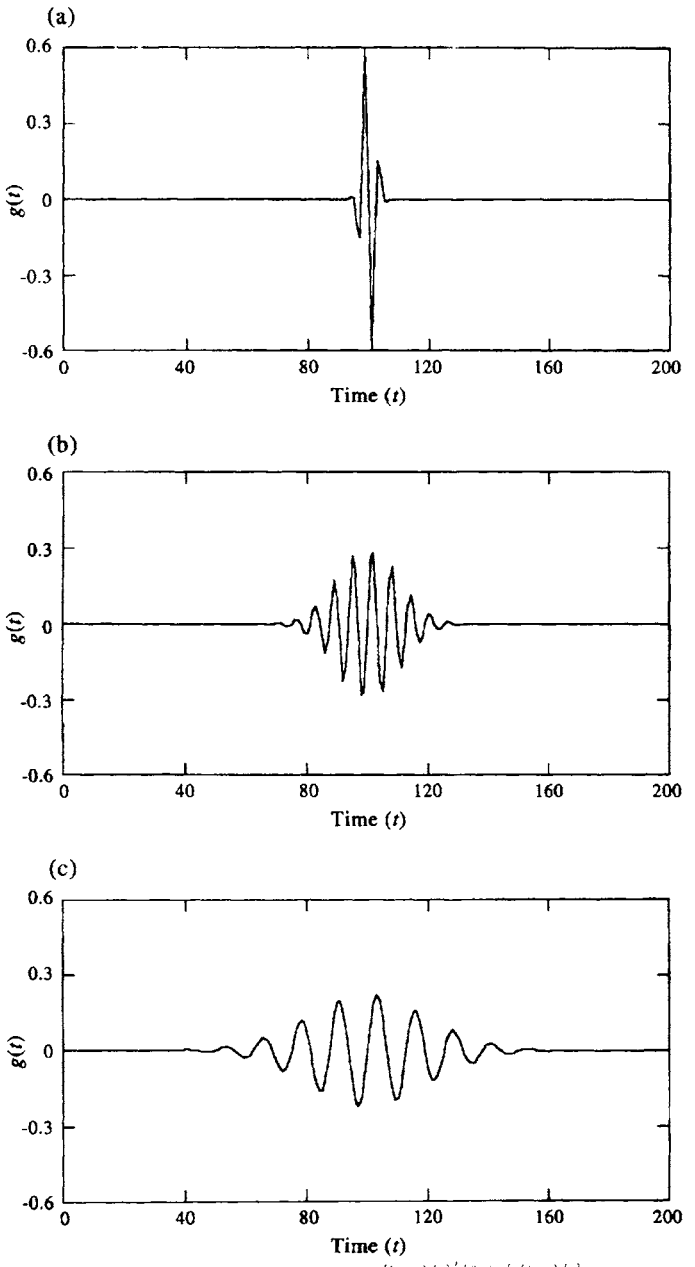


Figure 5.9.1. The Morlet wavelet, $g(t) = (1/\sqrt{a})e^{-[(t-\tau)/a]^2/2}\sin[c(t-\tau)/a]$, *where t is time in arbitrary units* ($t = t_n$; $n = 1, \ldots, 200$). *The example is for* $c = 10$ *and time lag* $\tau = 100$ *so that the wavelet is seen midway through the time series. (a)* $a = 2$; *(b)* $a = 10$; *(c)* $a = 20$.

sought in the time series. Large values of the transform $X_g(\tau, a)$ will then indicate where the time series $x(t)$ has the desired form. The simplest—and most time-consuming—method for obtaining the wavelet transform is to compute the transform at arbitrary points in parameter (τ, a) space using the discrete form of equation (5.9.2) for known values of $x(t)$ and $g(t)$. If one integrates from $0 < a \leq M$ and $0 < \tau \leq N$, the integration time goes as MN^2. An alternate method is to use the convolution theorem and then obtain the wavelet transform in spectral space

$$X_g[\tau, a] = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{i\tau\omega} G^*(a\omega) X(\omega)\, d\omega \tag{5.9.8}$$

where $G(\omega)$ and $X(\omega)$ are the Fourier transforms of $g(t)$ and $x(t)$, respectively. Since FFT transforms can now be exploited, the analysis time drops to $MN\log_2 N$. To use this method, $G(\omega)$ should be known analytically and the data must be preprocessed to avoid errors from the FFT algorithms. For example, if $x(t)$ is aperiodic, the discrete form of (5.9.7) will generate an artificial periodicity in the wavelet transform that greatly distorts the results for the end regions. Methods have been devised to work around this problem. Aliasing and bias in FFT routines must also be taken into account.

Meyers *et al.* (1993) used the standard Morlet wavelet (5.9.5), for which $g(t) = e^{-t^2/2} e^{ict}$, to examine a signal that changes frequency halfway through the measurement. Here, we have followed tradition and used c for frequency ω. After considerable attempts (including use of raw data, cosine weighted data and other variations), the authors decided that the best approach was to taper or buffer the original time series with added data points that attenuate smoothly to zero past the ends of the time series. "The region of the transform corresponding to these points is then discarded after the transform. Without this buffering, a signal whose properties are different near its ends will result in a wavelet transform that has been forced to periodicity at all scales through a distortion (in some cases severe) of the end regions. The greater the aperiodicity of the signal, the greater the distortion."

For the Morlet wavelet, the dilation parameter a giving the maximum correlation between the wavelet and a plane Fourier component of frequency ω_o (i.e. a wave of the form $e^{i\omega_o t}$) is

$$a_o = \frac{[c + (2 + c^2)^{1/2}]}{4\pi} T_o \tag{5.9.10}$$

where $T_o = 2\pi/\omega_o$ is the Fourier period. (In wavenumber space, T_o is replaced by wavelength λ_o and ω_o by k_o.) We note that any linear superposition of periodic components results in separate local maxima. Consequently, the wavelet transform of any function $x(t) = \sum A_j e^{ik_j t}$ will have modulus maxima at $a_j = [c + (2 + c^2)^{1/2}]/(2k_j)$.

5.9.3 Oceanographic examples

In this section, we will consider two oceanographic wavelet examples (surface gravity wave heights and zonal velocity from a satellite-tracked drifter) using the standard Morlet wavelet

$$g(t) \to g[(t - \tau)/a] = \frac{1}{\sqrt{a}} e^{-\frac{1}{2}[(t-\tau)/a]^2} \sin[c(t - \tau)/a] \tag{5.9.11}$$

In this real expression, the Gaussian function determines the envelope of the wavelet

while the sine function determines the wavelengths that will be preferentially weighted by the wavelet. The wavelet function progresses through the time series with increasing τ, its cone of influence centered at times $t = \tau$. As a increases, the width of the Gaussian spreads in time from its center value (Figure 5.9.1a–c). Increasing c increases the number of oscillations over the span of the function. The processing procedure is as follows: (1) read in the time series $x(n)$ ($n = 0, \ldots, N - 1$) to be analyzed, where $N = 2^m$ (m is an integer). To reduce ringing, extend each end of the time series by adding a trigonometric taper, tap $= 1 - \sin\phi$, where tap $= 1.0$ at the end values $x(0)$ and $x(N - 1)$. The total length of the buffered time series must remain a power of two; (2) remove the mean of the new record and then take the FFT of the time series to obtain $X(\omega)$; (3) take the Fourier transform of the wavelet $g(t)$ at given length scales, a, to obtain $G(a\omega)$; (4) calculate the integral (5.9.8) by convolving the product $G^*(a\omega)X(\omega)$ in Fourier space; (5) take the inverse FFT of the result to obtain $\sqrt{a}X_g[\tau, a]$ as a function of time dilation τ and amplitude, a.

In Figure 5.9.2(a) we have plotted a 300 s record of surface gravity wave heights measured off the west coast of Vancouver Island in the winter of 1993. Maximum wave amplitudes of around 3 m occurred mid-way through the time series. The Morlet wavelet transform of the record yields an estimate of the wave amplitude (Figure 5.9.2b) and phase (Figure 5.9.2c) as functions of the wave period (T) and time (t). Also plotted is the value of the wave period (T = scale a) at peak energy (Figure 5.9.2d). Comparison of Figures 5.9.2(b) and 5.9.2(d) reveals that the larger peaks near times of 75, 150, and 210 s all have about the same wavelet scale, a, corresponding to a peak wave period of around 8 s. Also, as one would expect, the 2π changes in phase between crests (Figure 5.9.2c) increases with increasing wave period (scale, a).

In our second example, we have applied a standard Morlet wavelet transform to a 90-day segment of 3-hourly sampled east–west (u) current velocity (Figure 5.9.3a) obtained from a satellite-tracked drifter launched in the northeast Pacific in August 1990 as part of the World Ocean Circulation Experiment (WOCE). The drifter was drogued at 15-m depth and its motion indicative of currents in the surface Ekman layer. The 90-day velocity record has been generated from positional data using a cubic spline interpolation algorithm. We focus our attention on the high-frequency end of the spectrum, $0 < a < 1.5$ days. As indicated by Figures 5.9.3(b) and (c), the first 30 days of the record, from Julian day (JD) 240 to 270, were dominated by weak semidiurnal tidal currents with periods of 0.5 days. Beginning on JD 270, strong wind-generated inertial motions with periods around 16 h ($f \approx 1.5$ cpd) dominated the spectrum. These energetic motions persisted through the record, except for a short hiatus near JD 295. A blow-up of the segment from JD 240 to 270 shows a rapid change in signal phase associated with the shift from semidiurnal tidal currents to near-inertial motions. The contribution from the beat frequency between the M_2 tidal signal and the inertial oscillations, $fM_2 = 0.0805 + 0.0621$ cph $= 0.1426$ cph can also be seen in the transformed data at period $T \approx 0.29$ days. Examination of the longer period motions ($2 < a < 30$ days) suggests the presence of a long-period modulation of the high-frequency motions associated with the near-inertial wave events.

5.9.4 The S-transformation

Wavelet transforms are not the only method for dealing with nonstationary oscillations with time-varying amplitudes and phases. The S-transformation (Stockwell *et al.*, 1994) is an extension of the wavelet transform that has been used by Chu (1994) to

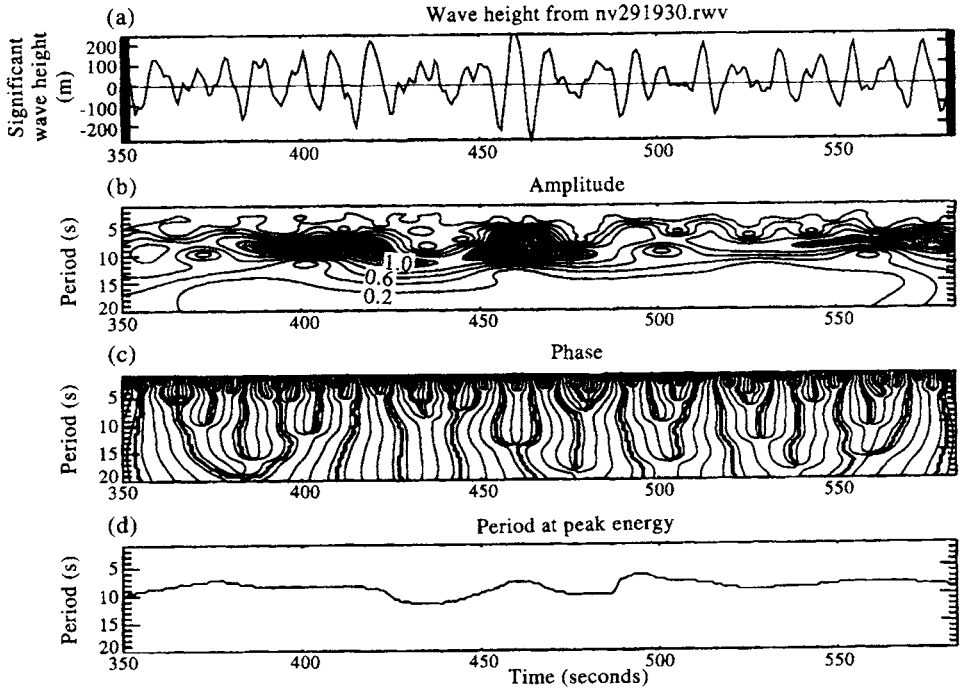


Figure 5.9.2. Morlet wavelet transform of surface gravity waves measured from a waverider buoy moored off the west coast of Vancouver Island. (a) Original five-minute time series of significant wave height for the winter of 1993. (b) Wave amplitude (m) and (c) phase (deg.) as a functions of time; (d) the value of a (wave period) at peak wave amplitude. (Courtesy, D. Masson.)

examine the localized spectrum of sea level in the TOGA data sets. For this particular transform, the relationship between the S-transform, $S(\omega, \tau)$, and the data, $x(t)$, is given by

$$S(\omega,\tau) = \int_{-\infty}^{\infty} H(\omega+\alpha)\mathrm{e}^{-(2\pi^2\alpha^2/\omega^2)}\mathrm{e}^{\mathrm{i}2\pi\alpha\tau}\mathrm{d}\alpha \tag{5.9.12}$$

$$x(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(\omega,\tau)\mathrm{e}^{\mathrm{i}2\pi\alpha\tau}\,\mathrm{d}\omega\,\mathrm{d}\tau \tag{5.9.13}$$

where

$$H(\omega+\alpha) = \int_{-\infty}^{\infty} x(t)\mathrm{e}^{-\mathrm{i}2\pi(\omega+\alpha)\tau}\,\mathrm{d}t \tag{5.9.14a}$$

$$= \int_{-\infty}^{\infty} S(\omega+\alpha,\tau)\,\mathrm{d}\tau \tag{5.9.14b}$$

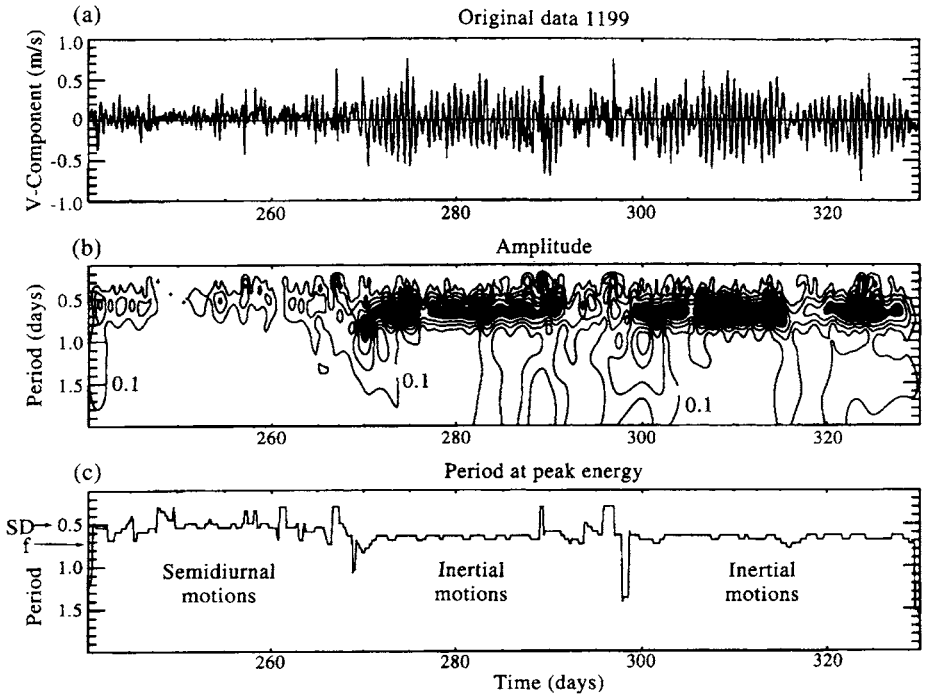


Figure 5.9.3. The Morlet wavelet transform of a 90-day record of the east–west velocity component from the trajectory of a satellite-tracked drifter in the northeast Pacific, September 1990. (a) Original 3-hourly time series; (b) amplitude (cm/s) versus time as a function of period, T, in the range $0 < T < 2.0$ days; (c) period (days) of the current oscillations at peak amplitude. (Courtesy, J. Eert.)

is the standard Fourier transform of the input time series data. As indicated by (5.9.14b), the Fourier transform is the time average of the S-transform, such that $|H(\omega)|^2$ provides a record-averaged value of the localized spectra $|S(\omega)|^2$ derived from the S-transform. Equation (5.9.13) can also be viewed as the decomposition of a time series $x(t)$ into sinusoidal oscillations which have time-varying amplitudes $S(\omega, \tau)$.

The discrete version of the S-transformation can be obtained as follows. As usual, let $x(t_n) = x(n\Delta t)$, $n = 0, 1, \ldots, N-1$ be a discrete time series of total duration $T = N\Delta t$. The discrete version of (5.9.12) is then

$$S(0, \tau_q) = \frac{1}{N}\sum_{m=0}^{n-1} x(m/T), \quad p = 0 \tag{5.9.15a}$$

$$S(\omega_p, \tau_q) = \sum_{m=0}^{N-1} \left\{ H[(m+p)/T] e^{-(2\pi^2 m^2/p^2)} e^{i2\pi mq/N} \right\}, \quad p \neq 0 \tag{5.9.15b}$$

where $S(0, \tau_q)$ is the mean value for the time series, $\omega_p = p/N\Delta t$ is the discrete frequency of the signal, and $\tau_q = q\Delta t$ is the time lag. The discrete Fourier transform is given by

$$H(p/T) = \frac{1}{N}\sum_{k=0}^{N-1} x(k/T) e^{-i2\pi pk/N} \tag{5.9.16}$$

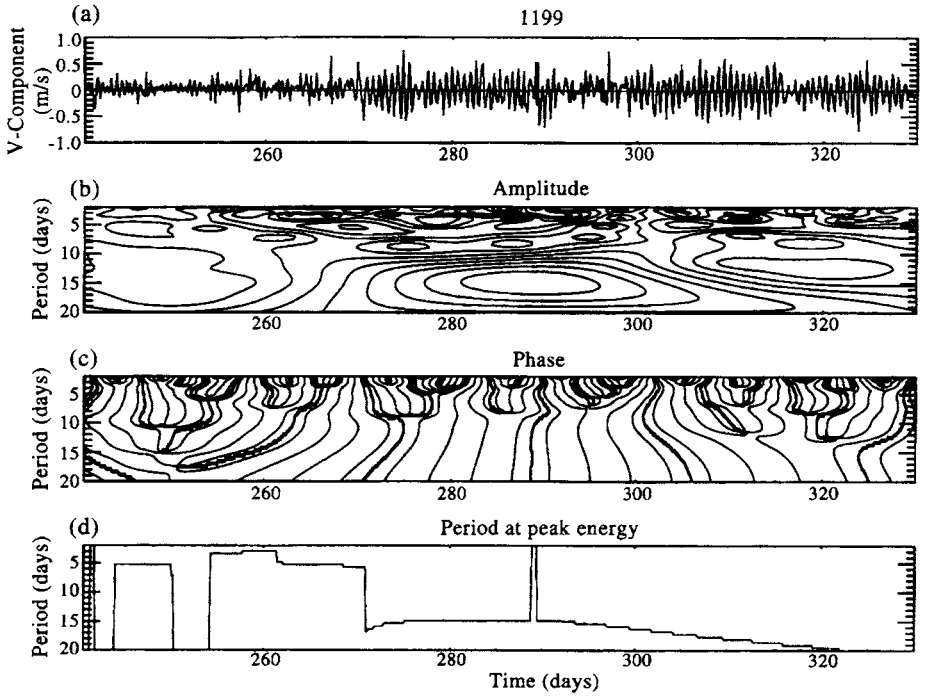


Figure 5.9.4. As for Figure 5.9.3 but for a larger range of periods. (a) Original 3-hourly velocity time series; (b) amplitude (cm/s) and (c) phase (degrees) versus time as a function of period, T, in the range $2 < T \leq 20$ days; (d) period (days) of the current oscillations at peak amplitude.

The S-transform is a complex function of frequency ω_p and time τ_q, with amplitude and phase defined by

$$A(\omega_p, \tau_q) = |S(\omega_p, \tau_q)| \tag{5.9.17a}$$

$$\Phi(\omega_p, \tau_q) = \tan^{-1}\{\mathrm{Im}[S(\omega_p, \tau_q)]/\mathrm{Re}[S(\omega_p, \tau_q)]\} \tag{5.9.17b}$$

For a sinusoidal function of the form

$$X(\omega_p, \tau) = A(\omega_p, \tau)\cos[2\pi\omega_p\tau + \Phi(\omega_p, \tau)] \tag{5.9.18}$$

the function X at frequency ω_p is called the "voice".

Chu (1994) applied the S-transform to the nondimensionalized sea-level records, $x(t)$, collected at Nauru (0°32′S, 166°54′W) in the western equatorial Pacific and La Libertad (2°12′S, 80°55′W) in the eastern equatorial Pacific. Here

$$x(t) = \frac{[\eta(t) - \bar{\eta}]}{\bar{\eta}} \tag{5.9.19}$$

and $\bar{\eta}(t)$ represents the mean value of the sea level, $\eta(t)$. A Fourier spectral analysis of the time series revealed a strong annual sea-level oscillation in the western Pacific and a weak annual oscillation in the eastern Pacific. Both stations had strong quasi-biennial oscillations with periods of 24–30 months. The S-transformation was then used to examine the temporal variability in these components throughout the 16 and

18-year time series. For example, the voices for the annual oscillation ($\omega_{16} = 16/T$; $T = 192$ months) were similar at the two locations with higher amplitudes in the late 1970s than in the late 1980s (Figure 5.9.5). At La Libertad, the annual cycle became weak after 1979. The temporally varying quasi-biennial oscillations ($\omega_8 = 8/T$) were out-of-phase between the western and eastern Pacific (Figure 5.9.6).

5.9.5 The multiple filter technique

The multiple filter technique is a form of signal demodulation that uses a set of narrow-band digital filters (windows) to examine variations in the amplitude and phase of dispersive signals as functions of time, t, and frequency, ω (or f). Originally designed to resolve complex transient seismic signals composed of several dominant frequencies (Dziewonski *et al.*, 1969), the technique has recently been modified for the analysis of clockwise and counterclockwise rotary velocity components (Thomson *et al.*, 1997) and in investigations of tsunami wave dispersion (Gonzalez and Kulikov, 1993).

The multiple filter technique relies on a series of band-pass filters centered on a range of narrow frequency bands to calculate the instantaneous signal amplitude or phase. Dziewonski *et al.* (1969) filter in the frequency domain rather than the time domain, although the results are equivalent to within small processing errors. The filtering algorithm generates a matrix (grid) of amplitudes or phases with columns representing time and rows representing frequency (or period). The gridded values can then be contoured to give a three-dimensional plot of the demodulated signal amplitude (or phase) as a function of time and frequency. Gonzalez and Kulikov (1993) used the technique to examine the evolution of tsunami waves generated by an undersea earthquake in the Gulf of Alaska on 6 March 1987 (Figure 5.9.7). Sea-level heights measured by two bottom-pressure recorders deployed in the deep ocean to the south of Kodiak Island show that the tsunami waves were highly dispersive (low frequencies propagated faster than high frequencies) and that the arrival times of the waves closely followed the theoretical predictions for shallow-water wave motions. Peak spectral amplitudes were centered around a period of roughly 5 min and the signal duration was about 40 min.

5.9.5.1 Theoretical considerations

Since the technique is used to examine signal energy as a function of time and frequency, it is desirable that the filtering function has good resolution in the immediate vicinity of each center frequency and time value of the f–t diagram. The Gaussian function was chosen to meet these requirements since the frequency–time resolution is greater for this function than any other type of nonband-limited function. A system of Gaussian filters with constant relative response leads to a constant resolution on a $\log(\omega)$ scale. If $\omega_n = 2\pi f_n$ denotes the center frequency of the nth row, the Gaussian window function can be written

$$H_n(\omega) = \exp\left\{-\alpha[(\omega - \omega_n)/\omega_n]^2\right\} \tag{5.9.20}$$

The Fourier transform of H_n, which bears a close resemblance to the Morlet wavelet (5.9.11), is

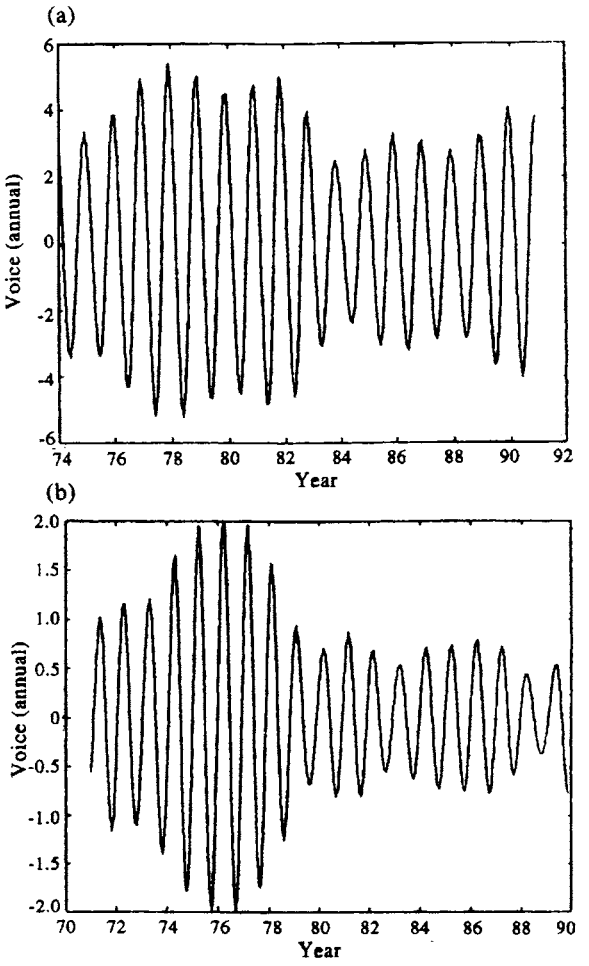


Figure 5.9.5. The "voices" for the annual oscillation ($\omega_{16} = 16/T$; $T = 192$ months) for (a) Nauru; and (b) La Libertad. Higher amplitudes were recorded in the late 1970s than in the late 1980s. (Chu, 1994.)

$$h_n(t) = \frac{\sqrt{\pi}}{2\alpha}\omega_n \exp\left[-(\omega_n^2 t^2/4\alpha)\right] \cos(\omega_n t) \tag{5.9.21}$$

The resolution is controlled by the parameter, α. The value of α that we choose depends on the dispersion characteristics in the original signal and, as the user of this method will soon discover, improved resolution in time means reduced resolution in frequency, and vice versa. We also need to truncate the filtering process. Dziewonski *et al.* (1969) used a filter cut-off where the filter amplitude was down 30 dB from the maximum.

If we let BAND be the relative bandwidth, then the respective lower and upper limits of the symmetrical filter, denoted $\omega_{L,n}$ and $\omega_{U,n}$, are

$$\omega_{L,n} = (1 - \text{BAND})\omega_n \tag{5.9.22a}$$

$$\omega_{U,n} = (1 + \text{BAND})\omega_n \tag{5.9.22b}$$

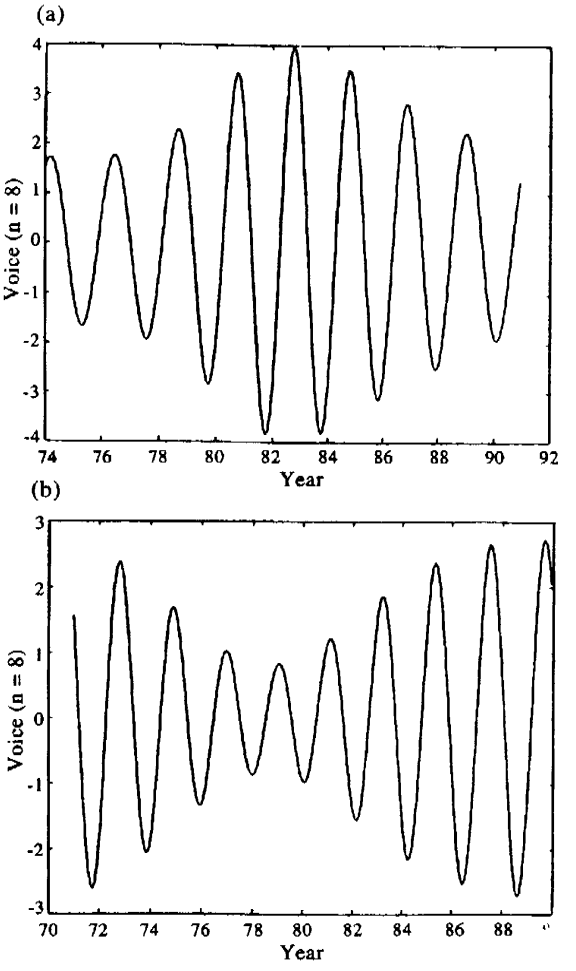


Figure 5.9.6. The "voices" for the quasi-biennial oscillations ($\omega_8 = 8/T$) for (a) Nauru; and (b) La Libertad. The oscillations were out-of-phase between the western and eastern Pacific. (Chu, 1994.)

The parameter α in (5.9.20) and (5.9.21) is expressed in terms of the bandwidth and the function β

$$\alpha = \beta/\text{BAND}^2 \tag{5.9.23}$$

where

$$\beta = \ln[H_n(\omega_n)/H_n(\omega_{L,n})] = \ln[H_n(\omega_n)/H_n(\omega_{U,n})] \tag{5.9.24}$$

describes the decay of the window function, $H_n(\omega)$. The window function then takes the form

$$H_n(\omega) = \begin{cases} 0 \quad \text{for} \quad \omega(1 - \text{BAND})\omega_n \\ \exp\{-\alpha[(\omega - \omega_n)/\omega_n]^2\} \quad \text{for } (1 + \text{BAND})\omega_n \leq \omega \leq (1 + \text{BAND})\omega_n \\ 0 \quad \text{for} \quad \omega > (1 + \text{BAND})\omega_n \end{cases} \tag{5.9.25}$$

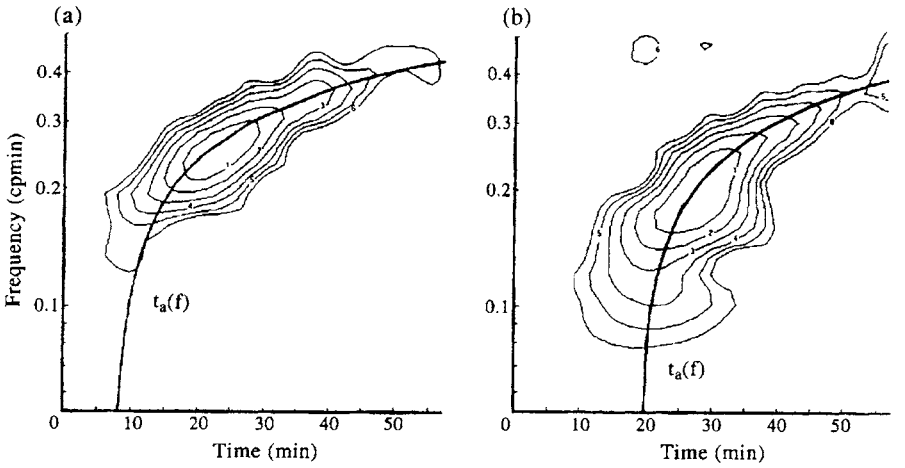


Figure 5.9.7. Multiple filter technique applied to sea-level heights measured in 5 km of water near 53°N, 156°W in the Gulf of Alaska on 6 March 1988. Amplitude contours in the f–t diagram are normalized by the maximum value and drawn with a step of 1 dB. Solid curve denotes the theoretical arrival time for these highly dispersive waves. (From Gonzalez and Kulikov, 1993.)

In their analysis of seismic waves, Dziewonski *et al.* (1969) used BAND = 0.25, $\beta = 3.15$, and $\alpha = \beta/\text{BAND}^2 = 50.3$.

The *f–t* diagram for the Alaska tsunamis (Figure 5.9.7) was obtained by windowing in the frequency domain with the truncated Gaussian function (5.9.25). In the time domain, the traces represent the convolution of the original data series with the Gaussian weighting function. The authors first set $\alpha = 25$ and chose $\beta = 1$, so that BAND = 0.20. The choice of β in (5.9.24) is arbitrary and can be set to unity, whereupon the bandwidth is determined by the e^{-1} values of the Gaussian function. For $\alpha = 25$ but $\beta = 2$, we have BAND = 0.28, and so on.

The flow chart for the analysis (Figure 5.9.8) is as follows:

(1) Remove the mean and trend (linear or other obvious functional trend) from the digital time series, $y(t)$.
(2) Fourier transform the time series. If an FFT algorithm is to used for this purpose, augment the time series with zeros to the nearest power of 2.
(3) Evaluate the center frequencies, $\omega_n = \omega_{n-1}/\text{BAND}$, for the array of narrow-band filters. The filters have a constant relative bandwidth, BAND, with the total width of each filter occupying the same number of rows in the log (frequency) scale. As noted on numerous occasions in the text, it is the length of the time series and the sampling rate which determine the frequency of the Fourier components. Since it often is difficult to get the frequencies obtained from the Fourier analysis to line up exactly with the center frequencies of the filters, select those components of the Fourier analysis which are closest to each member of the array and use these as the center frequencies.
(4) Select equally spaced times (columns) for calculation of amplitude or phase, focusing mainly on the times following the arrival of the waves.
(5) Filter the wave spectrum (sine and cosine functions of the Fourier transform) in the frequency domain with the Gaussian filter $H_n(\omega)$. This filter is symmetric about the center frequencies, ω_n.

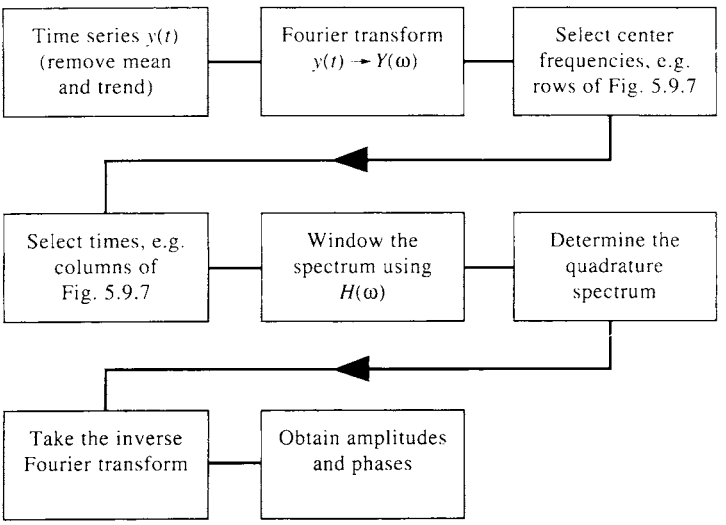


Figure 5.9.8. Flow chart for application of the multiple-filter technique. (Adapted from Dziewonski et al., 1969.)

(6) Take the inverse Fourier transform of the spectra using the same Fourier transform used in step 2. Since the inverse Fourier transform for the wave spectrum as windowed by the function $H_n(\omega)$ yields only the in-phase component of the filtered signal for each ω_n, knowledge of the quadrature spectrum is also required for evaluation of the instantaneous spectral amplitudes and phases. The quadrature spectrum is found from the in-phase spectrum using

$$Q_n(\omega) = H_n(\omega)e^{i\pi/2} \tag{5.9.26}$$

The amplitude and phase of the signal for each center frequency for each time are derived from the inverse Fourier transforms of the spectra and quadrature spectra.

(7) Instantaneous spectral amplitudes and phases are computed for each time step. The procedure (5)–(7) is repeated for each center frequency.

The multiple filter technique can be used to examine rotary components of current velocity fields. In this case, the input is not a real variable, as it is for scalar time series, but a complex input, $w(t) = u(t) + iv(t)$. Figure 5.9.9 is obtained from the analysis of a 90-day time series of surface currents measured by a 15 m drogued satellite-tracked drifter launched off the Kuril Islands in the western North Pacific on 4 September 1993 (Thomson *et al.*, 1997). The 3-hourly sampling interval used for this time series was made possible by the roughly eight position fixes per day by the satellite-tracking system. Plots show the variation in spectral amplitude of the clockwise and counterclockwise rotary velocity components as functions of time and frequency. For illustrative purposes, we have focused separately on the high and low frequency ends of the spectrum (periods shorter and longer than two days). Several interesting features quickly emerge from these f–t diagrams. For example, the motions are entirely dominated by the clockwise rotary component except within the narrow channel (Friza Strait) between the southern Kuril Islands where the motions become more rectilinear. The burst of clockwise rotary flow encountered by the drifter over the Kuril-Kamchatka Trench starting on day 28 was associated with wind-generated

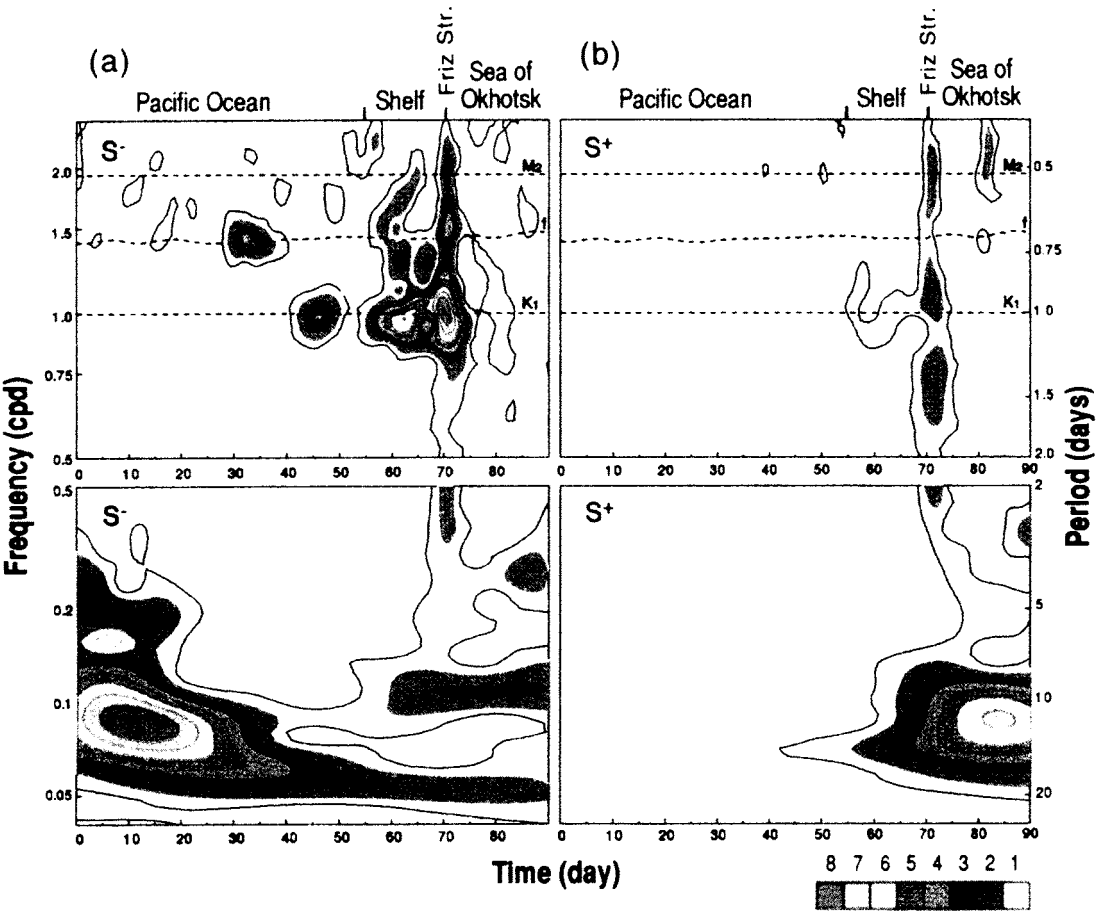


Figure 5.9.9. Multiple-filter technique applied to the velocity of a near-surface (15 m drogued) satellite-tracked drifter launched off the Kuril Island in 1993. S^- denotes the spectral amplitude (cm/s) of the clockwise rotary component versus frequency (cpd) and time (day); S^+ denotes the spectral amplitude of the counterclockwise component. (From Thomson et al., 1997.)

inertial waves whereas the strong clockwise rotary diurnal currents first encountered on day 40 and then again on day 55 were associated with diurnal-period continental shelf waves propagating along the steep continental slope of the Kuril Islands.

5.10 DIGITAL FILTERS

5.10.1 Introduction

Digital filtering is often an important step in the processing of digital oceanographic data. Applications include smoothing and decimation of time series, removal of fluctuations in selected frequency bands, and the alteration of signal phase. The term "decimation" originally meant the removal of every tenth point but is now commonly used for values other than 10. Digital filtering facilitates data processing by pre-conditioning the frequency content of the record. For example, filters are commonly used in studies of inertial waves to isolate current variability centered near the local Coriolis frequency, to remove background sea-level fluctuations in investigations of tsunamis, and to eliminate tidal frequency fluctuations in studies of low-frequency current oscillations (Figure 5.10.1). The terms "detided" or "residual" time series are

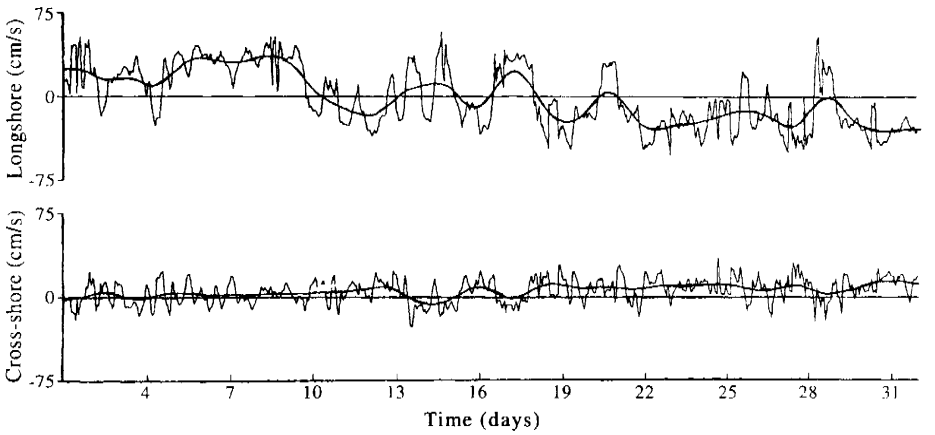


Figure 5.10.1. Time series of hourly longshore (top) and cross-shore (bottom) components of current velocity at 53-m depth on the continental shelf of northern Vancouver Island for March 1980. Thin line: original hourly data. Thick line: Hourly data filtered with a low-pass Godin tide-elimination filter, $A_{25}^2A_{24}/(25^224)$. (From Huggett, Crawford, Thomson, and Woodward, 1987.)

used to describe time series that have been filtered to remove tidal components. Filters also provide algorithms for data interpolation, for integration and differentiation of recorded signals, and for linear prediction models.

There is no single type of digital filter for general oceanographic use. Selection of an appropriate filter depends on a variety of factors, including the frequency content of the data and the kind of analysis to be performed on the filtered record. Personal preference or familiarity with one type of filter also can be deciding factors. However, in certain instances, one type of filter may be superior to another for a specific task, and proper filter selection involves some forethought. Often the type of filter must be tailored to the job at hand. For example, some of the so-called "tide-elimination" or "tide-killer" filters once used extensively in oceanography are inadequate for time series with marked diurnal period variability (Walters and Heston, 1982; Thompson, 1983). These filters permit leakage of unwanted diurnal tidal energy into the nontidal (residual) frequency bands of the filtered record. Elimination of this problem is possible through proper filter selection.

This chapter begins with a brief outline of basic filtering concepts then proceeds to descriptions of some of the more useful digital filters presently used in oceanographic research. We use the term "filter" to cover any linear operation on the data. In *optimal estimation* applications, the term applies specifically to an optimal estimate of the last measurement point. *Smoothing* is reserved for estimates spanned by the observations. Much of the emphasis in this chapter is on the design of low-pass digital filters that remove high-frequency oscillations from a given oceanographic time series. These filters can be used to construct other types of filters. The running-mean filter, the cosine Lanczos-window (or Lanczos-cosine) filter, and the Butterworth filter are among those most commonly used in oceanography.

5.10.2 Basic concepts

From a practical standpoint, a good low-pass filter should have five essential qualities: (1) a sharp cut-off, so that unwanted high-frequency components are effectively

removed; (2) a comparatively flat pass-band that leaves the low-frequency components unchanged; (3) a clean transient response, so that rapid changes in the signal do not result in spurious oscillations or "ringing" within the filtered record; (4) zero phase shift; and (5) acceptable computation time. As a rule, many of these desirable features are mutually exclusive and there are severe limitations to achieving the desired filter. We are invariably faced with a trade-off between the ability of the filter to produce the required results and the amount of filter-induced data loss we can afford to tolerate. For example, improved statistical reliability (increased degrees of freedom) for specified frequency bands decreases the frequency resolution of a filter while more sharply defined frequency cut-offs lead to greater ringing and associated data loss.

Suppose we have a time series consisting of the sequence

$$x(t_n) = x_n, \quad n = 0, 1, \ldots, N-1 \tag{5.10.1}$$

with observations at discrete times $t_n = t_o + n\Delta t$ in which t_o marks the start time of the record and Δt is the sampling increment. A digital filter is an algebraic process by which a sequential combination of the input $\{x_n\}$ is systematically converted into a sequential output $\{y_n\}$. In the case of linear filters, for which the output is linearly related to the input, the time domain transformation is accomplished through convolution (or "blending") of the input with the weighting function of the filter. Filters having the general form

$$y_n = \sum_{k=-M}^{M} h_k x_{n-k} + \sum_{j=-L}^{L} g_j y_{n-j}, \quad n = 0, 1, \ldots, N-1 \tag{5.10.2}$$

(in which M, L are integers and h_k, g_j are nonzero weighting functions) are classified as *recursive* filters since they generate the output by making use of a feed-back loop specified by the second summation term. Such filters "remember" the past in the sense that all past output values contribute to all future output values. Filters based on the input data only ($g_j = 0$), are classified as *nonrecursive* filters. Any filter for which $-M \leq k \leq M$ is said to be physically unrealizable (in the sense of any real-time output) because both past and future data are needed to calculate the output. Filters of this type have widespread application in the analysis of pre-recorded data for which all digital values are available beforehand. Filters for which $0 \leq k \leq M$ are said to be physically realizable or causal, and are used in real-time data acquisition and in forecasting procedures.

Impulse response: The output $\{y_n\}$ of a nonrecursive linear filter is obtained through the convolution

$$y_n = \sum_{k=-M}^{M} h_k x_{n-k} = \sum_{k=-M}^{M} h_{n-k} x_k, \quad n = 0, 1, \ldots, N-1 \tag{5.10.3}$$

where h_k are the time invariant weights and there are N data values $x_o, x_1, \ldots, x_{N-1}$. For a symmetric filter, the time domain convolution becomes

$$y_n = \sum_{k=0}^{M} h_k (x_{n-k} + x_{n+k}), \quad n = 0, 1, \ldots, N-1 \tag{5.10.4}$$

in which $h_k = h_{-k}$. The set of weights $\{h_k\}$ is known as the *impulse response* function and is the response of the filter to a spike-like impulse. To see this, we set $x_n = \delta_{0,n}$

where $\delta_{m,n}$ is the Kronecher delta function

$$\begin{aligned} \delta_{m,n} &= 0, \quad m \neq n \\ &= 1, \quad m = n \end{aligned} \tag{5.10.5}$$

Equation (5.10.1) then becomes

$$y_n = \sum_{k=-M}^{M} h_k \delta_{0,n-k} = h_n \tag{5.10.6}$$

The summations in equations (5.10.3) and (5.10.4) are based on a total of $2M + 1$ specified weights with individual values of h_k labeled by subscripts $k = -M, -M + 1, \ldots, M$. To make practical sense, the number of weights is limited to $M \ll N/2$ where $N\Delta t$ is the record length. In reality, it is not possible to use equation (5.10.3) to calculate an output value y_n for each time t_n. Because the response function spans a finite time (equal to $2M\Delta t$), difficulties arise near the ends of the data record and we are forced to accept the fact that there are always fewer output data values than input values. There are three options: (1) We can make do with $2M$ fewer estimates of y_n (resulting from time losses of $M\Delta t$ at each end of the record); (2) we can create values of $x(t_n)$ for times outside the observed range $0 \leq t < (N-1)\Delta t$ of the time series; or (3) we can progressively decrease the filter length, M, in accordance with the number of remaining input values. In the first approach, x_n is defined for $n = 0, 1, \ldots, N-1$ whereas y_n is defined for the shortened range $n = M, M+1, \ldots, N-(M+1)$. In the second approach, the appendaged estimates of x_n should qualitatively resemble the data at either end of the record. For example, we could use the "mirror images" of the data reflected at the end points of the original time series. In the third approach, the values y_{M-1} and $y_{N-(M-1)}$ are based on $(M-1)$ weights, the values y_{M-2} and $y_{N-(M-2)}$ on $(M-2)$ weights, and so on.

Frequency response: The Fourier transform of $y(t_n)$ in (5.10.3) is

$$\begin{aligned} Y(\omega) &= \sum_{n=-M}^{M} y_n e^{-i\omega_n \Delta t} \\ &= \sum_{k=-M}^{M} h_k e^{-i\omega_k \Delta t} \sum_{n=-M}^{M} x_{n-k} e^{-i\omega_{(n-k)} \Delta t} \\ &= H(\omega) X(\omega) \end{aligned} \tag{5.10.7}$$

so that convolution in the time domain corresponds to multiplication in the frequency domain. The function

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} = \sum_{k=-M}^{M} h_k e^{-i\omega k \Delta t}, \qquad \omega \equiv \omega_n = 2\pi n / N\Delta t \tag{5.10.8}$$

$n = 0, \ldots, N/2$ is known as the *frequency response* (or *admittance function*; see Section 5.8.7) since it determines how a specific Fourier component $X(\omega)$ is modified as it is transformed from input to output. For the symmetric filter (5.10.4), the transfer function reduces to

$$H(\omega) = h_0 + 2\sum_{k=1}^{M} h_k \cos(\omega k \Delta t) \tag{5.10.9}$$

Once $H(\omega)$ is specified, the weights h_k are found through the inverse Fourier transform

$$h_k = \sum_{n=-N/2}^{N/2} H(\omega) e^{i\omega_n k \Delta t} \tag{5.10.10}$$

In general, $H(\omega)$ is a complex function that can written in the form

$$H(\omega) = | H(\omega) | e^{i\phi(\omega)} \tag{5.10.11}$$

where the amplitude $| H(\omega) |$ is called the *gain* of the filter (a term originating with electrical circuitry) and $\phi(\omega)$ is the *phase lag* of the filter. The power $P(\omega)$ of the transfer function is given by

$$P(\omega) = H(\omega)H(-\omega) = H(\omega)H^*(\omega) = | H(\omega) |^2 \tag{5.10.12}$$

where, as usual, the asterisk denotes the complex conjugate.

5.10.3 Ideal filters

An ideal filter is one that has unity gain, $| H(\omega) | = 1$, at all frequencies within the specified *pass band(s)* and zero gain at frequencies within the *stop band(s)* (Figure 5.10.2). When processing oceanographic data, it is generally advantageous to have $\phi(\omega) = 0$ for all ω so that the filter produces no alteration in the phase of the frequency components. As we discuss in conjunction with recursive filters, zero phase shift can be guaranteed by first passing the input forward then backward (after inversion) through the same set of weights. In the case of nonrecursive filters, zero phase is accomplished using symmetric filters (i.e. those with no imaginary components).

Digital filters commonly used in processing oceanographic data can be classified under the general headings of *low-pass*, *high-pass*, or *band-pass* filters. Although impossible to achieve, we would like the amplitudes of our ideal filters to satisfy the following relations (see Figure 5.10.2)

$$\begin{aligned} \text{Low-pass: } | H(\omega) | &= 1 \text{ for } | \omega | \leq \omega_c \\ &= 0 \text{ for } \omega_c \leq \omega \end{aligned} \tag{5.10.12a}$$

$$\begin{aligned} \text{High-pass: } | H(\omega) | &= 0 \text{ for } | \omega | \leq \omega_c \\ &= 1 \text{ for } \omega_c \leq \omega \end{aligned} \tag{5.10.12b}$$

$$\begin{aligned} \text{Band-pass: } | H(\omega) | &= 1 \text{ for } \omega_{c1} \leq | \omega | \leq \omega_{c2} \\ &= 0 \text{ otherwise} \end{aligned} \tag{5.10.12c}$$

The *cut-off frequency*, $\omega_c (= 2\pi f_c)$, marks the transition from the pass-band to the stop-band. For ideal filters, the transition is step-like while for practical filters, the

transition has a finite width. In the latter case, ω_c is defined as the frequency at which the mean filter amplitude in the pass-band is decreased by a factor of $\sqrt{2}$ and should roughly coincide with spectral minima in the time series being analyzed; the power of the filter is down by a factor of 2 (−3 dB) at the cut-off frequency. As its name implies, a low-pass filter lets through (or is "transparent" to) low-frequency signals but strongly attenuates high-frequency signals (cf. Figures. 5.10.3a, b). High-pass filters let through the high-frequency components and strongly attenuate the low-frequency components (cf. Figures. 5.10.3a, c). Band-pass filters permit only frequencies in a limited range (or band) to pass unattenuated.

Low-pass filters are the most common filters used in oceanographic data analysis. It is through these filters that low-frequency, long-term variability of oceanographic signals is determined. The running-mean filter, which involves a moving average over an odd number of values, is the simplest form of low-pass filter. More complex filters with better frequency responses, such as the low-pass Kaiser–Bessel window used in Figure 5.10.3(b), also are commonly used. High-pass filtered data are readily obtained by subtracting the low-pass filtered data from the original record from which the low-pass data were derived. One does not need to create a separate high-pass filter. Similarly, band-pass filters can be formed by an appropriate combination of low-pass and high-pass filters. In the ocean, seawater acts as a form of natural low-pass filter, attenuating high-frequency wave or acoustic energy at a much more rapid rate than low-frequency energy. Acoustic waves of a few hertz (cycles per second) can propagate thousands of kilometers in the ocean whereas acoustic waves of hundreds of kilohertz are strongly attenuated over a few hundred meters.

High-pass filters are less frequently used than low-pass filters. Applications include the delineation of high-frequency, high-wavenumber fluctuations in the internal wave band (roughly $2f < \omega < N$, where N is the Brunt–Väisälä frequency) and the isolation of seiche or tsunami motions in closed or semi-enclosed basins. Band-pass filters are used to isolate variability in relatively narrow frequency ranges such as the near-inertial frequency band or, in North America, the electronic-induced 60-cycle noise in high-frequency oceanic data caused by AC power supplies.

The maximum range of frequencies that can be covered by a digital filter is determined at the high-frequency end by the Nyquist frequency, $\omega_N = \pi/\Delta t$ (radians/unit time), and at the low-frequency end by the fundamental frequency, $\omega_1 = 2\pi/T$, where $T = N\Delta t$ is the length of the record. The corresponding range in cycles/unit time are determined by $f_N = 1/(2\Delta t)$ and $f_1 = 1/T$. Provided that the cut-off frequencies are sufficiently far removed from the ends of the intervals, digital filters can be applied throughout the range, $\omega_1 < |\omega| < \omega_N$ ($f_1 < |f| < f_N$).

5.10.3.1 Bandwidth

The difference in frequency between the two ends of a pass-band defines an important property known as the *bandwidth* of the filter. To illustrate the relevance of this property, we consider an ideal band-pass filter with constant gain, linear phase, and cut-off frequencies ω_{c1}, ω_{c2} such that

$$\begin{aligned} H(\omega) &= H_o \exp(-i\omega t_o), \quad \omega_{c1} \leq |\omega| < \omega_{c2} \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{5.10.13}$$

From (5.10.12c), the impulse response is

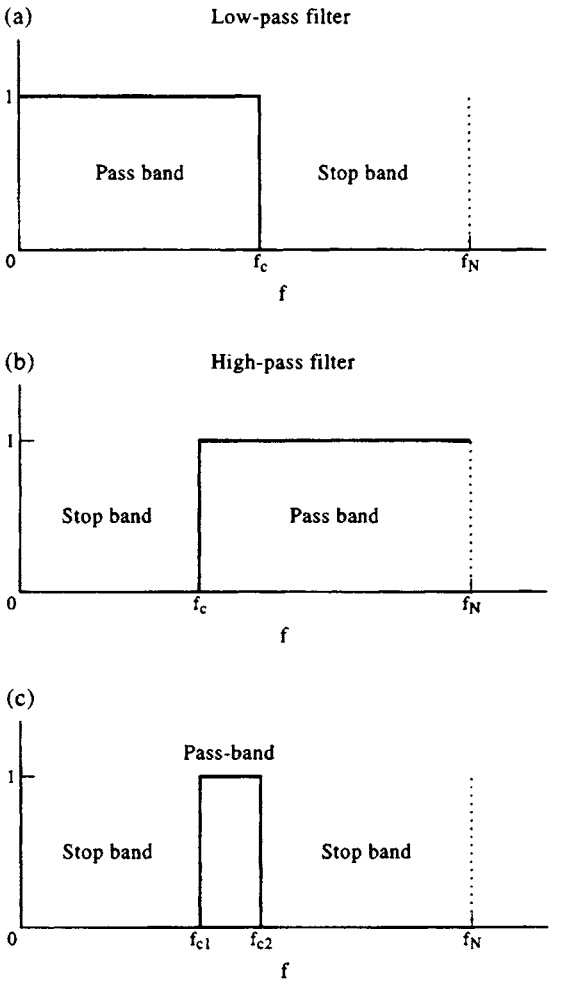


Figure 5.10.2. Frequency response functions, $|H(f)|$, for ideal filters. (a) Low pass; (b) high pass; and (c) band pass. The band-pass filter has been constructed from the combined low-pass and high-pass filters. f_N and f_c are the Nyquist and cut-off frequencies, respectively.

$$h_k = \frac{1}{2\pi} H_o \left(\int_{\omega_{c1}}^{\omega_{c2}} e^{-i\omega t_o} e^{i\omega k\Delta t} \, d\omega + \int_{\omega_{c1}}^{\omega_{c2}} e^{i\omega t_o} e^{-i\omega k\Delta t} \, d\omega \right)$$
$$= \frac{2H_o}{\pi} \Delta\omega \cos\left[\Omega(k\Delta t - t_o)\right] \frac{\sin\left[\Delta\omega(k\Delta t - t_o)\right]}{\Delta\omega(k\Delta t - t_o)} \tag{5.10.14}$$

in which $\Omega = \frac{1}{2}(\omega_{c1} + \omega_{c2})$ is the center frequency and $\Delta\omega = \omega_{c2} - \omega_{c1}$ is the bandwidth. For high or low-pass filters, the bandwidth is equal to the cut-off frequency.

Using the fact that $\sin p/p \to 1$ as $p \to 0$, we find that the peak amplitude response of the filter (5.10.14) is directly proportional to the bandwidth $\Delta\omega$ as $\Delta\omega(k\Delta t - t_o) \to 0$. Note also that a narrow-band filter (one for which $\Delta\omega \to 0$) will

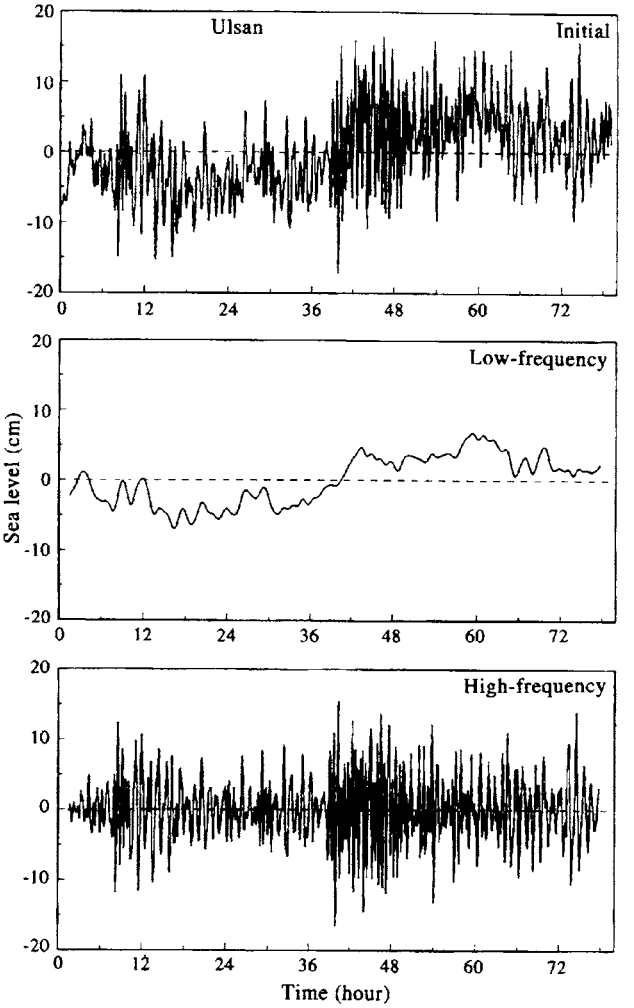


Figure 5.10.3. Filtering of a tide gauge record for Ulsan, Korea using low and high-pass Kaiser–Bessel filters (windows) with length $T/27 = 3$ h; $T = 81$ h is the record length and $\Delta t = 0.5$ min the sampling increment. (a) Original record; (b) low-pass filtered record; (c) high-pass filtered record. (Courtesy, A. Rabinovich.)

oscillate longer (i.e. persist to higher values of k) than a broad-band filter when subjected to a transient loading. Put another way, the persistence of the ringing that follows the application of the filter to a data set increases as the bandwidth decreases. From a practical point of view, this means that the ability of a filter to resolve sequential transient events is inversely proportional to the bandwidth. The narrower the bandwidth (i.e. the finer the resolution in frequency), the longer the time series needed to resolve individual events. For example, if we use a band-pass filter to isolate inertial frequency motions in the range 0.050–0.070 cph, the bandwidth $\Delta f = \Delta\omega/2\pi = 0.020$ cph and the filter could accurately resolve inertial events that occurred about $1/\Delta f = 50$ h apart. If we now reduce the bandwidth to 0.010 cph, the filter is only capable of resolving transient motions that occur more than 100 h apart.

(The need to have long records to resolve closely spaced frequencies is exactly the problem we faced in Section 5.6.5.4 regarding the Rayleigh criterion for tidal analysis.)

Another way of stating the above relationship is that the uncertainty in frequency, Δf(or $\Delta\omega$), is inversely proportional to the length of time T over which the signal oscillates (i.e.$\Delta f \approx 1/T$) so that $T\Delta f \approx 1$ for a given filter. If we wish to use a filter with a very narrow bandwidth, we need to analyze long time-series records in which the signals of interest, such as the tides, have a high degree of persistence. In terms of observed data, the measured bandwidth of an oscillation in current speed, sea-level elevation, or other oceanic parameter is directly related to the persistence of the signal. For example, a wind-generated clockwise rotary inertial current having an observed bandwidth $\Delta f \approx 0.10$ cpd implies that the burst of inertial energy had a duration $T \approx 1/\Delta f = 10$ days.

5.10.3.2 Gibbs' phenomenon

In practice, step-like transfer functions such as described by equation (5.10.12) are not possible. Digital filters invariably possess finite-slope transition zones between the stop and pass-bands. To illustrate some of the fundamental impediments to creating ideal filters, consider the step-like transfer function

$$\begin{aligned} H(\omega) &= 1 \qquad 0 < \omega \leq \omega_N \\ &= 0 \qquad -\omega_N \leq \omega < 0 \end{aligned} \tag{5.10.15}$$

(Figure 5.10.2a) where, for convenience, we specify a cut-off frequency $\omega_c = 0$. Assuming that $H(\omega)$ is repeated over multiples of the basic interval $(-\omega_N, \omega_N)$, the appropriate Fourier series expansion for equation (5.10.15) is given in the usual manner by

$$H(\omega) = \frac{1}{2}a_o + \sum_{n=1}^{\infty} [a_n \cos(\omega n \Delta t) + b_n \sin(\omega n \Delta t)] \tag{5.10.16}$$

with coefficients

$$a_n = \frac{1}{\omega_N} \int_{-\omega_N}^{\omega_N} H(\omega) \cos(\omega n \Delta t)\, d\omega \tag{5.10.17a}$$

$$b_n = \frac{1}{\omega_N} \int_{-\omega_N}^{\omega_N} H(\omega) \sin(\omega n \Delta t)\, d\omega \tag{5.10.17b}$$

The fact that $a_n = 1$, for all n, suggests reformulation of the problem in terms of the function

$$H_c(\omega) = H(\omega) - 1/2 \tag{5.10.18}$$

centered about $H(\omega) = 1/2$, the mean functional value at the discontinuity. Since $H(\omega)$ is then an odd function, cosine terms in (5.10.16) can be eliminated immediately. Moreover, H_c is symmetric about $\omega = \pm\frac{1}{2}\omega_N = \pm\pi/(2\Delta t)$ so that there are no even sine terms. For odd n, (5.10.17b) yields $b_n = 2/n\pi$ and (5.10.16) becomes

$$H(\omega) = \frac{1}{2} + \frac{2}{\pi}\left[\sin(\omega\Delta t) + \frac{\sin(3\omega\Delta t)}{3} + \frac{\sin(5\omega\Delta t)}{5} + \ldots\right] \tag{5.10.19}$$

which must be truncated after a finite number of terms.

Successive approximations to the series (5.10.19), and hence to the function (5.10.15), are not convergent near discontinuities such as that for the step-like transition region of the ideal high-pass filter shown in Figure 5.10.4. In this example, the filter amplitude $|H(\omega)|$ is zero for $\omega < \omega_c$ (the stop band) and unity for $\omega_c < \omega < \omega_N$ (the pass band). The succession of overshoot ripples, or ringing, is known as *Gibbs' phenomenon*. The ripple period, $T = p\pi t$ (p is an integer), is fixed but increasing the number of terms in the Fourier series for $H(\omega)$ decreases the distortion due to the overshoot effects. However, even in the limit of infinitely many terms, Gibbs' phenomenon persists as the amplitude of the first overshoot diminishes asymptotically to about 0.18 or about 9% of the pass-band amplitude. The first minimum decreases asymptotically to about 5% of the pass-band amplitude. In the limit of large $N \to \infty$, it can be shown (Godin, 1972; Hamming, 1977) that

$$H_\infty(0) \to \frac{1}{\pi}\int_0^\pi \sin u/u \, du \tag{5.10.20}$$

The values of $H_\infty(0)$ can be found in tables of the sine integral function. In the case of Figure 5.10.4, the value for the first maximum is 1.08949 (= 1.0 + 0.08949) while that for the first undershoot is 0.9514 (= 1.0 − 0.04858).

Gibbs' phenomenon has considerable importance in that it occurs whenever a function has a discontinuity. For example, suppose that we want to use equation (5.10.19) to remove spectral components near a cut-off frequency, ω_c. Unless the spectral components in the stop and pass-bands are well separated relative to the width of the transition zone, the finite ripples will cause leakage of unwanted energy into the filtered record. Noise from the stop-band will not be completely removed and certain frequencies in the pass-band will be distorted. A critical aspect of filter design is the attenuation of the overshoot ripples using smoothing or tapering functions (windows). As discussed in Section 5.6.6, windows are important in reducing side-lobe leakage in spectral estimates.

Further difficulties arise when we apply the weights $\{h_k\}$ of an ideal filter in the time domain. Consider the nonrecursive, low-pass filter (positive frequency only)

$$\begin{aligned} H(\omega) &= 1, \quad 0 \le \omega \le \omega_c \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{5.10.21}$$

for which the impulse function is, for $k = -N, \ldots, N$

$$\begin{aligned} h(t_k) = h_k &= \frac{1}{\omega_N}\sum_{\omega=0}^{\omega_c} \cos(\omega k\Delta t)\Delta\omega \\ &= \frac{\sin(\omega_c k\Delta t)}{\omega_N k\Delta t} \\ &= \frac{f_c}{f_N}\frac{\sin(2\pi f_c k\Delta t)}{2\pi f_c k\Delta t} \end{aligned} \tag{5.10.22}$$

in which $h_o = f_c/f_N$. The weights h_k attenuate slowly, as $1/k$, so that a large number of terms are needed if the filter response $H(\omega)$ is to be effectively carried over to the time domain. In addition to being computationally inefficient, filters constructed from a large number of weights lead to considerable loss of information at the ends of the data sequence. Practical considerations force us to truncate the set of weights thereby enhancing the overshoot problem associated with Gibbs' phenomenon in the frequency domain. Moreover, if we truncate the length of the data set (5.10.1), we are unable to accurately replicate (5.10.21) in the frequency domain. This leads to a finite slope between the stop and pass-bands of the filter.

The situation is similar for high-pass filters

$$\begin{aligned} H(\omega) &= 0, \quad 0 \leq \omega \leq \omega_c \\ &= 1, \quad \text{otherwise} \end{aligned} \tag{5.10.23a}$$

In this case

$$\begin{aligned} h_k &= \frac{1}{\omega_N} \sum_{\omega=\omega_c}^{\omega_N} \cos(\omega k \Delta t) \Delta\omega \\ &= -\frac{f_c}{f_N} \frac{\sin(2\pi f_c k \Delta t)}{2\pi f_c k \Delta t}, \quad k = -N, \ldots, N \end{aligned} \tag{5.10.23b}$$

where $h_o = 1 - f_c/f_o$. Notice that, except for the central term h_o, the weights h_k of the high-pass filter (5.10.23b) are equal to minus the weights h_k of the low-pass filter

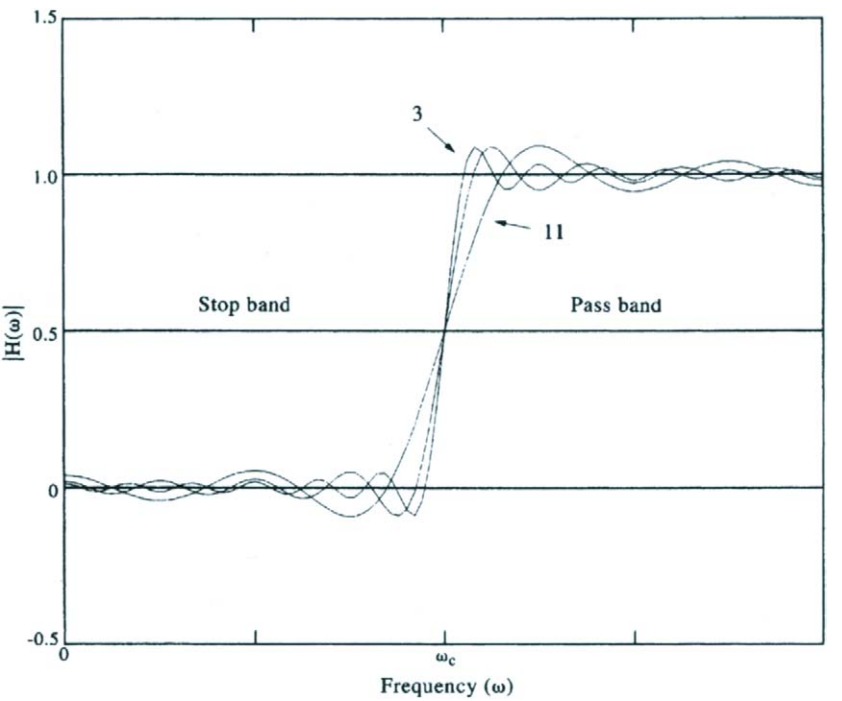


Figure 5.10.4. Gibbs' phenomenon (overshoot ripples) arising from successive approximations to the step-like function $|H(\omega)| = 1$, $\omega_c < \omega \leq \omega_N$, and zero otherwise. $\omega_c = 2\pi f_c$ is the cut-off frequency. Curves are derived from (5.10.19) using $M = 3$, 7, and 11 terms.

(5.10.22). The center value, h_o of the high-pass filter is found from h_o of the low pass filter by: $h_o(high\ pass) = 1 - h_o\ (low\ pass)$.

The difficulties that arise with Gibbs' phenomenon are somewhat alleviated by applying smoothing functions that attenuate the overshoot ripples. As usual, the price we pay for improved decay of the weighting terms is a broadening of the main lobe centered at the frequency being filtered. As we remarked earlier, the fact that the transition from the pass to the stop-band takes place over a finite range of frequencies necessitates a working definition for the cut-off frequency, ω_c. Here, ω_c is defined as the frequency at which the power $|H(\omega)|^2$ of the filter is attenuated by a factor of 2 (−3 db) from its mean pass-band value. (Power in dB = $20 \log(A/A_o)$ where A_o is a reference level for the signal amplitude, A, having power proportional to A^2.) Alternatively, the cut-off frequency marks the frequency at which the amplitude $|H(\omega)|$ of the filter is reduced by a factor of $\sqrt{2}$ of the pass-band amplitude (amplitude in dB = $10 \log(A/A_o)$).

5.10.3.3 Recoloring

The transfer function amplitude $|H(\omega)|$ defines the effectiveness of a particular filter in transmitting or blocking power within specific frequency bands. Since no filter is perfect, in the sense that its transfer function is exactly unity throughout the pass band(s) and zero in the stop band(s), it is often necessary to "re-color" (re-scale) the output $Y(\omega)$ so that the total variance in the pass-band spectral estimates equals the total variance of the input data for that frequency range. The need to re-color stems from practical considerations involving the choice of filter, cut-off frequency, and filter steepness through the transition band. For a pass-band of width $\Delta\omega$, multiplication of the filter output $|Y(\omega)|$ by a frequency-independent correction factor γ given by

$$\gamma(\Delta\omega) = \frac{\text{input variance within bandwidth}}{\text{output variance within bandwidth}}$$

ensures that the output power is adequately re-scaled.

We can illustrate the re-coloring process using the Hanning (von Hann) and Hamming windows. If $x(t)$ is any scalar time series of length N, and $y(t)$ is the filtered output of this series following application of one of these windows, then the Fourier transform of the output, $Y(f_k)$, for discrete frequencies $f_k = (k/T), k = 0, 1, \ldots, (N/2)$ is given by

$$Y(f_k) = 0.50X(f_k) - 0.25X(f_{k-1}) - 0.25X(f_{k+1}) \quad \text{(Hanning)} \tag{5.10.24a}$$

$$Y(f_k) = 0.54X(f_k) - 0.23X(f_{k-1}) - 0.23X(f_{k+1}) \quad \text{(Hamming)} \tag{5.10.24b}$$

where $X(f_k)$ is the Fourier transform of the original time series. The corresponding expected values for $|Y(f_k)|^2$ in (5.10.24a, b) are

$$E[|Y(f_k)|^2] = (0.50)^2 + (0.25)^2 + (0.25)^2 = 0.3750 \tag{5.10.24c}$$

$$E[|Y(f_k)|^2] = (0.54)^2 + (0.23)^2 + (0.23)^2 = 0.3974 \tag{5.10.24d}$$

so that the spectral density estimates $S(f_k) \approx |Y(f_k)|^2]$ for each frequency component of a time series smoothed by a Hanning window should be rescaled by the exact factor

$(0.375)^{-1} = 8/3$ to correct for the loss of power due to the filter. For the Hamming window, the factor is roughly $(0.397)^{-1} \approx 5/2$. Note that, according to equation (5.10.24), we can easily obtain spectral estimates $S(f_k)$ for each windowed time series by summing up the squared amplitudes $|X(f)|^2$] of three adjacent Fourier components of the original time series

$$S(f_k) = C_o|X(f_k)|^2 + C_{-1}|X(f_{k-1})|^2 + C_{+}1|X(f_{k+1})|^2 \tag{5.10.25}$$

where $C_o = 0.50$ and $C_{-1} = C_{+1} = -0.25$ for the Hanning window and $C_o = 0.54$ and $C_{-1} = C_{+1} = -0.23$ for the Hamming window.

5.10.4 Design of oceanographic filters

The isolation of signal variability within specific frequency bands requires filters with well-defined frequency characteristics. The design of application-specific filters can proceed in two basic ways. The first approach is to assemble a combination of simple filters, such as moving averages of variable length, and from them construct a filter with the required characteristics. This is referred to as *cascading* since the output from the lead-off filter is used as input to the second filter, output from the second filter is used as input to the third, and so on. Filter cascading is used in the design of Godin's (1972) tide-elimination filters and the squared Butterworth filters described later in this section. The second approach is to specify the desired characteristics of the filter precisely and then use poles and zeros of mathematical functions to design a filter that meets these requirements as closely as possible. As an example, we might wish to eliminate the annual cycle from a long time-series of upper-ocean variability, such as sea surface temperature, so that weaker fluctuations are no longer overwhelmed by the dominant seasonal changes. The filter properties are then directly tailored to the processing requirements and to the data specific to the region of interest. (In this example, we could also use least-squares analysis to determine the annual cycle and then subtract this cycle from the original data.)

Regardless of which approach is taken, it is important that the impulse and frequency response functions of the filter have a number of fundamental properties: (1) The frequency response function should have reasonably sharp transitions between adjacent stop and pass bands, especially if the data do not have wide "spectral-gaps" between dominant frequencies within the two bands. At the same time, the transition should not be so steep as to introduce large side-lobe effects or cause the filter output to become unstable; (2) the transfer function should have nearly constant amplitude and zero phase (even symmetry) within the pass and stop bands so that corrections to amplitude and phase are easily applied. Linear phase change as a function of frequency is acceptable but requires corrective work at the end of the processing; and (3) the impulse response should have as short a span as possible to both minimize the number of points lost (or that are need to be appended at the ends of the data) and to reduce the amount of computation.

5.10.4.1 Frequency versus time domain filtering

In most instances, filters are designed to precondition the frequency content of the data prior to further analysis. This immediately suggests that the design of a filter begin with specification of the transfer function, $H(\omega)$. Once $H(\omega)$ has been

determined there are two ways to proceed. The standard time-domain approach (e.g. Hamming, 1977) is to Fourier transform $H(\omega)$ to obtain the time-domain filter weights, h_k, which are then used in the convolution (5.10.3) to determine the output $\{y_n\}$. The output is subsequently Fourier transformed to determine $Y(\omega)$. The frequency domain approach (e.g. Walters and Heston, 1982; Middleton, 1983) makes use of the fact that $Y(\omega) = H(\omega)X(\omega)$, where $X(\omega)$ is the Fourier transform of the data $x(t)$. In this approach, the data are Fourier transformed to obtain $X(\omega_i)$. $i = 1, 2, ..., N/2$, where $X(\omega)$ consists of a set of $N/2$ frequency-dependent amplitudes and phases $[A(\omega_i), \phi(\omega_i)]$ at discrete frequencies. The filtered record is obtained by multiplying $X(\omega)$ by $H(\omega)$. The time-domain series $\{y_n\}$ can be derived from the inverse Fourier transform of $Y(\omega)$.

There are pros and cons for both approaches. The time-domain approach uses the actual recorded data and filtering consists of simple sums and products. Moreover, the filtered series $\{y_n\}$ can be immediately plotted against the original input $\{x_n\}$ to see directly the effectiveness of the filter. Discontinuities in the time series, which lead to transient filter ringing effects, can be dealt with on the spot. However, if the calulation of $Y(\omega)$ and its associated spectral estimate $| Y(\omega) |^2$ are the ultimate goals, the time-domain approach requires application of two Fourier transforms: First, we use $H(\omega)$ to define the filter weights $\{h_k\}$ and then transform $y_n \rightarrow Y(\omega)$ to obtain the Fourier components. This can lead to roundoff and computational errors.

In the frequency-domain analysis, only one Fourier transform, $x_n \rightarrow X(\omega)$, is required. On this basis, it seems preferable to use the Fourier transform method and just set to zero all those frequency components outside the range of interest. The filtered data $\{y_n\}$ are then found through an inverse transform of the modified Fourier components, $Y(\omega) = H(\omega)X(\omega)$. One obvious difficulty with this procedure is that the discrete frequencies of the Fourier estimates may not be properly positioned relative to the required cut-off frequency of the filter; that is, the cut-off frequency may fall mid-way between two discrete Fourier components. Walters and Heston (1982) also pointed out that the sharp cut-off associated with this process causes ringing through the entire data set (Figure 5.10.5). For this reason, the Fourier coefficients must be reduced gradually to zero over a range of frequencies. For example, Nowlin *et al.* (1986) used a trapezoidal-shaped band-pass filter to study inertial oscillations in data collected in Drake Passage. In this particular instance, "Fourier coefficients within 0.03 cpd of the local inertial frequency were retained undiminished, and this central portion was flanked by two tapered sections 0.06-cpd wide in which the coefficients were reduced linearly to zero." The smooth filter transition results in a substantial reduction in ringing in the filtered data but is certainly reminiscent of data tapering required in the time-domain analysis. A more detailed discussion of frequency domain filtering is presented in Section 5.10.9.

5.10.4.2 Filter cascades

In some instances, a desired filter $H(\omega)$ can be constructed from a series or *cascade* of basis filters $H_j(\omega)$ such that

$$H(\omega) = H_1(\omega) \times H_2(\omega) \times ... \times H_q(\omega) \tag{5.10.26}$$

where "$\times$" denotes successive applications of individual transfer functions, beginning with H_1. That is, the data are first processed with $H_1(\omega)$ and the output from this filter

passed through $H_2(\omega)$; the output from $H_2(\omega)$ is then passed through $H_3(\omega)$, and so on until the last filter, $H_q(\omega)$. The final output from $H_q(\omega)$ corresponds to the sought-after output from $H(\omega)$. Although the technique is straightforward and helps to minimize roundoff error, it has a number of major drawbacks, including the need for extended computations and the possibility of repeated ringing as one filter after another is applied in succession.

A high-pass filter $H_H(\omega)$ is obtained from its low-pass counterpart $H_L(\omega)$ by the relation $H_H(\omega) = 1 - H_L(\omega)$ where, in theory, the combined output from the two filters simply recreates the original data, since $H_L(\omega) + H_H(\omega) = 1$. This has advantages in situations where $H_L(\omega)$ is easily derived or is already available. In the time domain, the high-pass filtered record $\{y'_n\}$ is obtained by subtracting the output $\{y_n\}$ from the low-pass filter form the input time series $\{x_n\}$. Care is needed to ensure that the times of y_n and x_n are properly aligned so that $y'_n = x_n - y_n, n = M$, $M + 1, \ldots, N - 2M$.

A band-pass filter can be constructed from an appropriate high and low-pass filter using the method illustrated in Figure 5.10.2(c). Here, the cut-off frequency of the low-pass filter becomes the high-frequency cut-off of the band-pass filter; similarly, the cut-off frequency of the high-pass filter becomes the low-frequency cut-off of the band-pass filter. The cascade then has the form $H_B(\omega) = H_L(\omega) \times H_H(\omega)$.

Because nonrecursive filters are symmetric ($H(\omega)$ is a real function), there is no shift in phase between the input and output signals. This feature of the filters, as well as their general mathematical simplicity, has contributed to their popularity in oceanography. Recursive filters, on the other hand, are typically nonsymmetric. This introduces a frequency-dependent phase shift between the input and output variables and adds to the complexity of these filters for oceanic applications. Despite these difficulties, recursive filters are useful additions to any processing repertoire. The good news is that we can remove phase shifts introduced through the "forward" application of the filter by reversing the process and passing the data "backward"

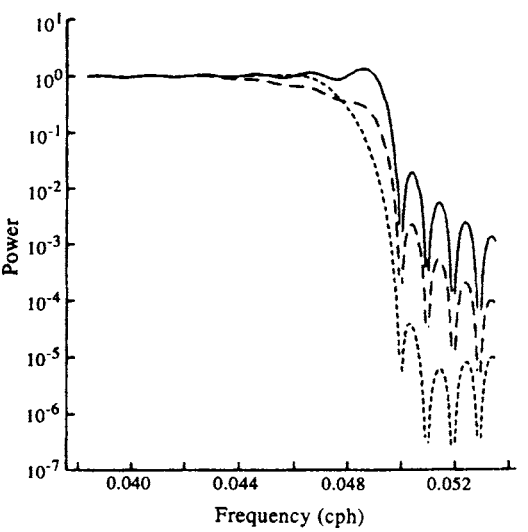


Figure 5.10.5. Frequency-response functions for low-pass filters with different transition bands. Solid line: A step-like transition band. Long-dashed line: A nine-point cosine-tapered transition band. Short-dashed line: A three-point optimally designed transition band. The cut-off associated with each filter causes ringing through the entire data set. (From Elgar, 1988.)

through the filter. In performing the latter step, we must be careful to invert the order of the record values between the forward and backward passes. Specifically, if the recursive filter introduces a phase shift $\phi(\omega)$ at frequency ω (or equivalently, a time shift $\phi/\omega = \phi/2\pi f$), it will introduce a compensating shift $-\phi(\omega)$ when passed in the reverse order through the filter. To show this sequence let $x_1, x_2, \ldots, x_n$ be the original data sequence used as input to a given filter with nonzero phase characteristics, and $y_1, y_2, \ldots, y_n$ the output from the filter (Figure 5.10.6). If we now invert the order of the output and pass the inverted signal through the filter again, we obtain a new output z_1, $z_2, \ldots, z_n$. The order of the z-output is then inverted to form $z_n, z_{n-1}, \ldots, z_1$, which returns us to the proper time sequence. For simplicity we can rewrite this later sequence as $y'_1, y'_2, \ldots, y'_n$. The act of applying the filter a second time cancels any phase change from the first pass through the filter. Note that this corresponds to squaring the transfer function so that the final transfer function for the recursive filter is $|H(\omega)|^2$.

As an example of a phase-dependent recursive filter, consider the high-pass *quasi-difference filter*

$$y(n\Delta t) = x(n\Delta t) - \alpha x[(n-1)\Delta t] \tag{5.10.27a}$$

where α is a parameter in the range $0 < \alpha \leq 1; \alpha = 1$ corresponds to the simple difference filter (Koopmans, 1974). The transfer function for this filter is

$$H(\omega) = 1 - \alpha e^{-i\omega\Delta t} \tag{5.10.27b}$$

and the phase function is

$$\phi(\omega) = \tan^{-1}[\alpha \sin(\omega\Delta t)/(1 - \alpha\cos(\omega\Delta t))] \tag{5.10.27c}$$

Reversing the order of the output from the first pass of the data through the filter and then running the time-inverted record through the filter again is tantamount to passing the data through a second filter $H(\omega)^*$. This introduces a phase change $-\phi(\omega)$ which cancels the phase change $\phi(\omega)$ from the first filter (Figure 5.10.7). The symmetric filter obtained from this cascade is then

$$\begin{aligned} |H(\omega)|^2 &= H(\omega) \times H(\omega)^* \\ &= (1 - \alpha e^{-i\omega\Delta t})(1 - \alpha e^{+i\omega\Delta t}) = [1 - 2\alpha\cos(\omega\Delta t) + \alpha^2]^{1/2} \end{aligned} \tag{5.10.27d}$$

5.10.5 Running-mean filters

The *running-mean* or *moving-average filter* is the simplest and one of the most commonly used low-pass filters in physical oceanography. In a typical application, the filter (which is simply a moving rectangular window) consists of an odd number of $2M + 1$ equal weights, h_k, $k = 0, \pm 1, \ldots, \pm M$, having constant values

$$h_k = \frac{1}{2M+1} \tag{5.10.28a}$$

where h_k resembles a uniform probability density function in which all occurrences are equally likely. The running-mean filter produces zero phase alteration since it is symmetric about $k = 0$, it satisfies the normalization requirement

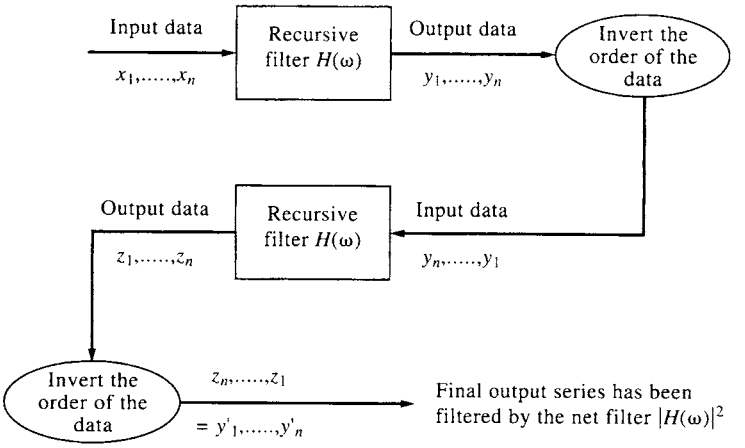


Figure 5.10.6. The processing sequence for a nonsymmetric recursive filter $H(\omega)$ which removes phase changes $\phi(\omega)$ introduced to the data sequence x_i ($i = 1, \ldots, n$) by the filter. This cascade produces a symmetric squared-filter response $|H(\omega)|^2$.

$$\sum_{k=-M}^{M} h_k = 1 \tag{5.10.28b}$$

and is straightforward to apply. To obtain the output sequence $\{y_m\}$ for input sequence $\{x_n\}$, the first $2M + 1$ values of x_n (namely $x_o, x_1, \ldots, x_{2M}$) are summed and then divided by $2M + 1$, yielding the first filtered value $y_M = y(2M\Delta t/2)$. The subscript M reminds us that the filtered value replaces the original data record x_M at the appropriate location in the time series. The next value, $y_M + 1$, is obtained by advancing the filter weights one time step Δt and repeating the process over the data sequence $x_1, x_2, \ldots, x_{2M+1}$ and so on up to $N - 2M$ output values. The $\{y_m\}$ consist of a "smoothed" data sequence with the degree of smoothing, and associated loss of information from the ends of the input, dependent on the number of filter weights. Mathematically

$$y_{M+i} = \frac{1}{2M+1} \sum_{j=0}^{2M} x_{i+j}, \quad i = 0, \ldots, N - 2M \tag{5.10.29}$$

A high-pass running-mean filter can be generated by subtracting the output $\{y_m\}$ from the original data. The output $\{y'_m\}$ for the high-pass filter is

$$y'_m = x_m - y_m, \quad m = M, M + 1, \ldots, N - 2M \tag{5.10.30}$$

where we make certain we subtract data values for the correct times. This technique of obtaining a high-pass filtered record from a low-pass filtered record will also be applied to other types of filters.

The transfer function $H(\omega)$ for the running-mean filter is given by equation (5.10.8). Using equation (5.10.27) and the fact that $\Delta t = \pi/\omega_N$, we find that

$$H(\omega) = \frac{1}{2M+1} \left\{ \frac{1 + 2\sin[(\pi/2M)(\omega/\omega_N)]\cos[\pi/2(M+1)(\omega/\omega_N)]}{\sin(\pi/2\,\omega/\omega_N)} \right\} \tag{5.10.31a}$$

$$= \frac{1}{2M+1} \frac{\sin[\pi/2(2M+1)(\omega/\omega_N)]}{\sin[(\pi/2)(\omega/\omega_N)]} \tag{5.10.31b}$$

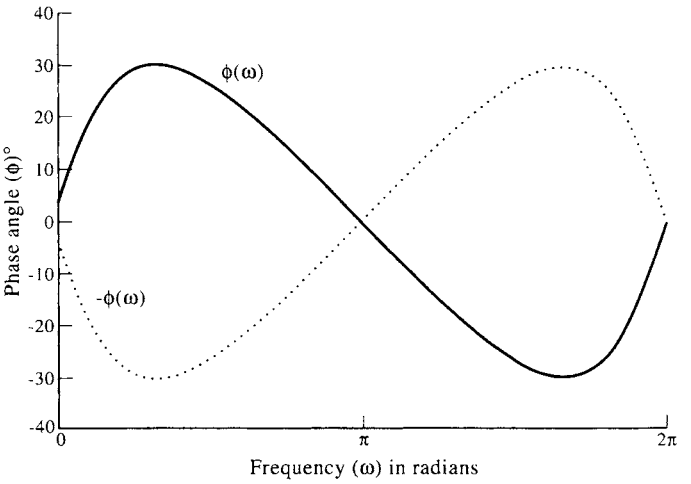


Figure 5.10.7. The phase change $\phi(\omega)$ for a quasi-difference filter (with $\alpha = 0.5$ as a function of frequency, ω).

where $H(\omega) \to 1$ as $\omega/\omega_N \to 0$. As M increases, the central lobe of the transfer function narrows (Figure 5.10.8) and the cut-off frequency (at which $|H(\omega)| = e^{-1}|H(0)|$) moves closer to zero frequency. The filter increasingly isolates the true mean of the signal. Unfortunately, the filter has considerable contamination in the stop-band due to the large, slowly attenuating side-lobes. Reduction of these side-lobe effects requires a long filter which means severe loss of data at either end of the time series. The running-mean filter should therefore only be used with long data sets ("long" compared with the length of the filter). Accurate filtering requires use of more sophisticated filters.

For the three-point weighted average, $h_k = 1/3$ and equation (5.10.31) yields

$$\begin{aligned} H(\omega; 3) &= \frac{1}{3}[1 + 2\cos(\pi\omega/\omega_N)] \\ &= \frac{1}{3}\frac{\sin[(3\pi/2)(\omega/\omega_N)]}{\sin[(\pi/2)(\omega/\omega_N)]} \end{aligned} \tag{5.10.32}$$

while for five-point weighted average, $h_k = 1/5$ and

$$H(\omega; 5) = \frac{1}{5}\frac{\sin[(5\pi/2)(\omega/\omega_N)]}{\sin[(\pi/2)(\omega/\omega_N)]} \tag{5.10.33}$$

(Figure 5.10.8). Numerous examples of running-mean filters appear in the oceanographic literature. A common use of running-mean filters is to convert data sampled at times t to an integer multiple of this time increment for use in standard analysis packages. Data collected at intervals Δt of 5, 10, 15, 20, or 30 min are usually converted to hourly data for use in tidal harmonic programs, although the least-squares algorithms used in these programs also work with unequally spaced time-series data (e.g. Foreman, 1977, 1978). Running-mean filters also are commonly used to create weekly, monthly, or annual time series (Figure 5.10.9).

5.10.6 Godin-type filters

For the low-pass filtering of sub-hourly sampled tidal records prior to decimation to "standard" hourly values, Godin (1972) recommends the use of cascaded running-mean filters with response functions of the form

$$\frac{A_n^2 A_{n+1}}{n^2(n+1)}, \qquad \frac{A_n A_{n+1}^2}{n(n+1)^2} \tag{5.10.34}$$

Here, A_n and A_{n+1} are the average values of n and $n + 1$ consecutive data points, respectively. Each filter smoothes the data three times. In the first version in (5.10.34), the smoothing is performed twice using the n-point average and once using the $\{n + 1\}$-point average. The alternative version uses the $\{n + 1\}$-point average twice and the n-point average once. Following the filter operation, the smoothed records can then be sub-sampled at hourly intervals without concern for aliasing by higher-frequency components. For the second version in (5.10.34), the response function is

$$H(\omega) = \frac{1}{n^2(n+1)} \frac{\sin^2[(\pi/2)(n\omega/\omega_N)] \sin[(\pi/2)(n+1)\omega/\omega_N]}{\sin^3[(\pi/2)(\omega/\omega_N)]} \tag{5.10.35}$$

Godin filters $(A_{12}^2 A_{14})/(12^2 14)$ are used routinely to smooth oceanographic time series sampled at multiples of 5-min increments prior to their use in tidal analysis programs. On the other hand, 30-min data would first be smoothed using the filter $(A_2^2 A_3)/(2^2 3)$ (Figure 5.10.10) and then decimated to hourly data. Conversion of 30-min data from Aanderaa RCM4 current meters to hourly data requires such a three-stage running-average filter. The filter is needed to convert the instantaneous directions and average speeds from the current meter to quantities more closely resembling vector-averaged currents. Application of the moving low-pass filter (5.10.34) removes high-frequency components and helps avoid the aliasing errors that would occur if the raw data were simply decimated to hourly values without any form of prior smoothing. Simply picking out a value each hour is, of course, akin to not having recorded the higher frequency variability in the first place. Some care is required in that the smoothing process reduces the amplitude of various Fourier components outside the tidal band. As a result, amplitudes of Fourier components derived after application of the filter must be corrected (recolored) in inverse proportion to the amplitude of the filter at the particular frequency. Phases of the Fourier components are unaltered by this symmetric filter.

The formulation (5.10.34) also can be used to generate low-pass filters to remove diurnal, semidiurnal, and shorter period fluctuations from the hourly records. Although these filters have been criticized in recent years because of their slow transition through the high-frequency end of the "weather band" (periods longer than two days), they are easy to apply, have good response in the daily tidal band and consume relatively little data from the ends of the time series. The most commonly used version of the low-pass Godin filter is $(A_{24}^2 A_{25})/(24^2 25)$ in which the hourly data are smoothed twice using the 24-point (24-h) average and once using the 25-point average. The filter response is

$$\begin{aligned} H(\omega) &= \frac{1}{24^2 25} \sin^2[24(\pi/2)(\omega/\omega_N)] \frac{\sin[25(\pi/2)(\omega/\omega_N)]}{\sin[(\pi/2)(\omega/\omega_N)]} \\ &= \frac{1}{24^2 25} \sin^2(24\pi f \Delta t) \frac{\sin(25\pi f \Delta t)}{\sin^3(\pi f \Delta t)} \end{aligned} \tag{5.10.36}$$

where as before $\omega = 2\pi f$ (f is in cycles per hour), $\omega_N = \pi/\Delta t$ and $\Delta t = 1$ h. Note that a

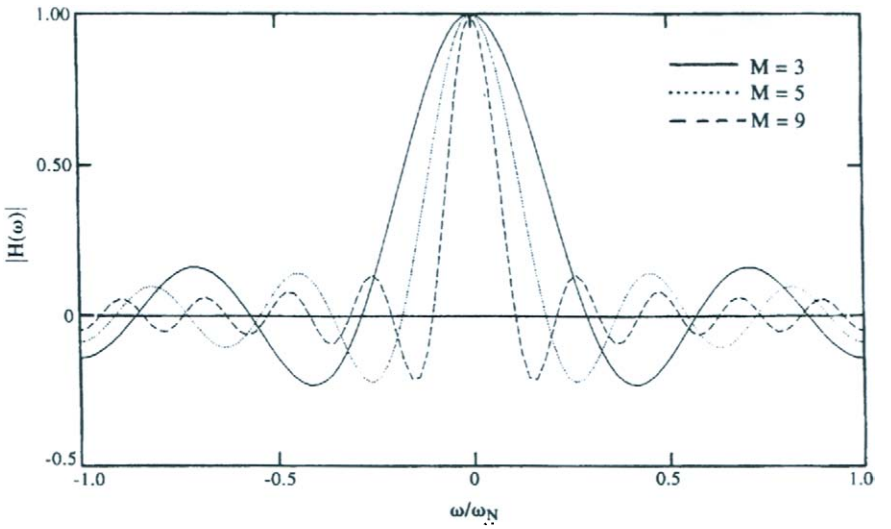


Figure 5.10.8. The frequency response functions, $|H(\omega)|$, for running-mean (weighted average) filters for $M = 3, 5, 9$. ω_N = Nyquist frequency.

total of 35 data points (i.e. 35 h) are lost from each end of the time series and that the filter has a half-amplitude point near 67 h (Figure 5.10.11). The weights of this symmetric 71-h-length filter are (Thompson, 1983)

$$\begin{aligned} h_k &= \frac{1/2}{24^2 25}[1200 - (12-k)(13-k) - (12+k)(13+k)], \quad 0 \le k \le 11 \\ &= \frac{1/2}{24^2 25}(36-k)(37-k), \quad 12 \le k \le 35 \end{aligned} \tag{5.10.37}$$

The Godin low-pass filter (5.10.37) effectively removes all daily tidal period energy except for slight leakage in the diurnal frequency band. More precisely, the filter eliminates variability due to the principal mixed diurnal constituent, K_1, for which the amplitude is down by 3.2×10^{-3}, and is only slightly less effective in removing variability due to the declinational diurnal constituent, O_1. The filter represents a marked improvement over the simple A_{24} and A_{25} running-mean filters and Doodson filter commonly used earlier for tidal analysis (cf. Groves, 1955). The principal failing of the Godin filter is its relatively slow transition between the pass and stop-bands which leads to significant attenuation of nontidal variability in the range of two to three days. This shortcoming of the filter has inspired a number of authors to investigate more efficient techniques for removing the high-frequency portion of oceanographic signals. The cosine–Lanczos filter, the transform filter, and the Butterworth filter are often preferred to the Godin filter, or earlier Doodson filter, because of their superior ability to remove tidal period variability from oceanic signals.

5.10.7 Lanczos-window cosine filters

As mentioned in Section 5.10.3.2, transfer functions for ideal (rectangular) filters are formulated in terms of truncated Fourier series. This leads to overshoot ripples (Gibbs' phenomenon) near the cut-off frequency with subsequent leakage of unwanted signal energy into the pass band. *Lanczos-window cosine filters* are reformulated rectangular filters which incorporate a multiplicative factor (the *Lanczos window*) to

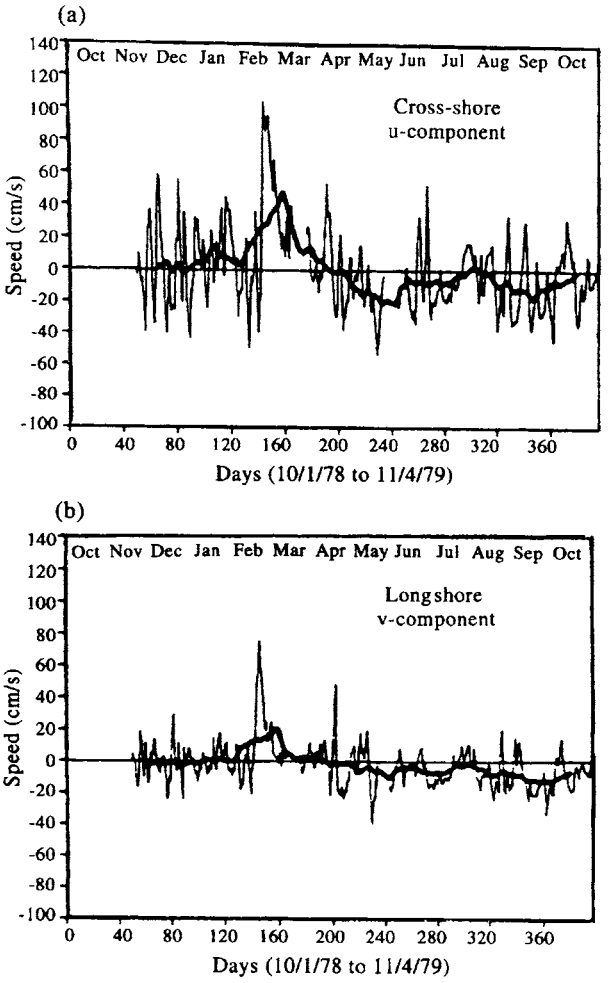


Figure 5.10.9. Daily mean time series of cross-shelf (top) and longshelf (bottom) near-surface currents off Cape Romain in the South Atlantic Bight for the period 10 January 1979 to 11 April 1979. Thin line: Daily average data. Thick line: 30-day running-mean values. (From McClain et al., 1988.)

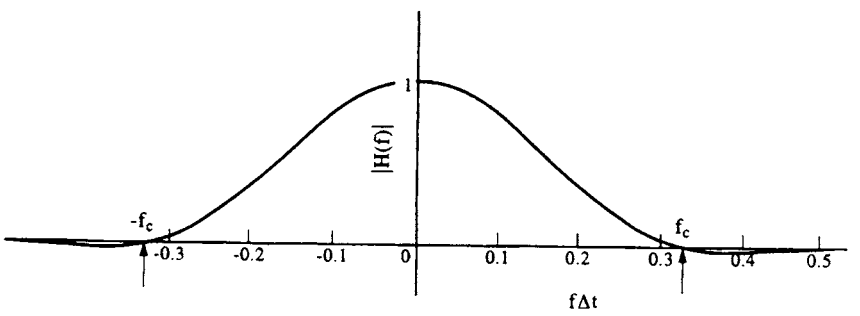


Figure 5.10.10. The frequency-response function, $|H(f)|$, for the Godin-type filter $A_2^2A_3/(2^23)$ used to smooth 30-min data to hourly values. The horizontal axis has units $f\Delta t$, with $f_N\Delta t = 0.5$; f_c is the cut-off frequency. (From Godin, 1972.)

rectangular filters which incorporate a multiplicative factor (the *Lanczos window*) to ensure more rapid attenuation of the overshoot ripples. A variety of other windows can also be used. The terms *Lanczos–cosine filter* and *cosine–Lanczos filter* are commonly used names for a family of filters using windows to reduce the side-lobe ripples. Owing to their simplicity and favorable characteristics, these filters have gained considerable popularity among physical oceanographers over the years (Mooers and Smith, 1967; Bryden, 1979; Freeland *et al.*, 1986).

5.10.7.1 Cosine filters

We start with an ideal, low-pass filter with transfer function

$$\begin{aligned} H(\omega) &= 1 \quad 0 \leq |\,\omega\,| \leq \omega_c \\ &= 0 \quad \text{elsewhere} \end{aligned} \tag{5.10.38}$$

and assume that the function $H(\omega)$ is periodic over multiples of the Nyquist frequency domain $(-\omega_N, \omega_N)$. Written as Fourier series, the response function is

$$H(\omega) = \frac{a_o}{2} + \sum_{k=1}^{M} [a_k \cos(\omega k \Delta t) + b_k \sin(\omega k \Delta t)] \tag{5.10.39}$$

where we have truncated the series at $M \ll N$; as usual, N is the number of data points to be processed by the filter. To eliminate any frequency-dependent phase shift, we insist that $H(\omega) = H(-\omega)$, whereby $b_k = 0$. The resulting *cosine filter* has the transfer function

$$H(\omega) = h_o + \sum_{k=1}^{M} h_k \cos(\pi k \omega / \omega_N) \tag{5.10.40}$$

where coefficients h_k $(= \frac{1}{2}a_k)$ given by

$$h_k = \frac{1}{\omega_N} \int_o^{\omega_N} H(\omega) \cos(\pi k \omega / \omega_N)\, d\omega \tag{5.10.41}$$

with $k = 0, 1, \ldots, M$. The weighting terms h_k are those which determine the output series $\{y_n\}$ for given $\{x_n\}$. We assume that M is sufficiently large that $H(\omega)$ is close to unity in the pass-band and near zero in the stop-band.

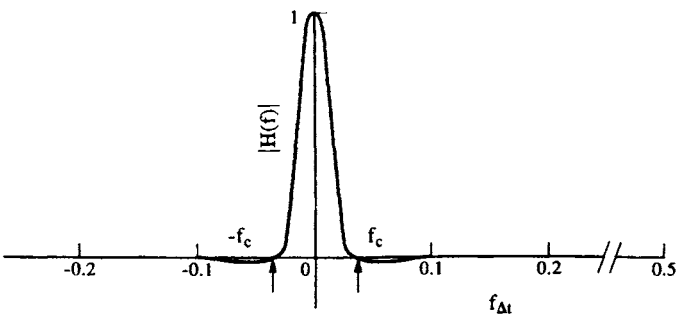


Figure 5.10.11. Same as Figure 5.10.10 but for the Godin-type low-pass filter $A_{25}^2 A_{24}/(25^2 24)$ used to eliminate tidal oscillations in hourly data. (From Godin, 1972.)

For a low-pass cosine filter, $0 \leq |\omega| \leq \omega_c$ defines the bounds of the integral (5.10.41) and the weights are given by

$$h_k = \frac{\omega_c}{\omega_N} \frac{\sin(\pi k \omega_c/\omega_N)}{\pi k \omega_c/\omega_N}, \quad k = 0, \pm 1, \ldots, \pm M \tag{5.10.42}$$

for which $h_o = \omega_c/\omega_N$. The corresponding weights for a high-pass filter, $|\omega| > \omega_c$, are

$$h_o = 1 - \omega_c/\omega_N, \quad k = 0 \tag{5.10.43}$$

$$h_k = \frac{-\omega_c}{\omega_N} \frac{\sin(\pi k \omega_c/\omega_N)}{\pi k \omega_c/\omega_N}, \quad k = \pm 1, \ldots, \pm M \tag{5.10.44}$$

That is, h_o (*high pass*) $= 1 - h_o$ (*low pass*) while for $k \neq 0$, the coefficients h_k are simply of opposite sign. The functions (5.10.42) and (5.10.44) are identical to those discussed in context of Gibbs' phenomenon. Thus, the cosine filter is a poor choice for accurately modifying the frequency content of a given record based on preselected stop and pass-bands. As an example of the response of this filter, Figure 5.10.12 presents the transfer function

$$H(\omega) = 0.4 + 2 \sum_{k=1}^{9} [\sin(0.4k\pi)/k\pi] \cos(k\omega)$$

for a low-pass cosine filter with $\omega_c/\omega_N = 0.4$ and $M = 10$ terms. This filter response is compared to the ideal low-pass filter response and to the modified cosine filter using the Lanczos window (with sigma factors) discussed in the next section.

5.10.7.2 The Lanczos window

Lanczos (1956) showed that the unwanted side-lobe oscillations of the form $\sin(p)/p$ in equations (5.10.42) and (5.10.44) could be made to attenuate more rapidly through use of a smoothing function or window. The window consists of a set of weights that successively average the (constant period) side-lobe fluctuations over one cycle, with the averaging period determined by the last term kept or the first term ignored in the Fourier expansion (5.10.44). In essence, the window acts as a low-pass filter of the weights of the cosine filter. The Lanczos window is defined in terms of the so-called *sigma-factors* (cf. Hamming, 1977)

$$\sigma(M,k) = \frac{\sin(\pi k/M)}{\pi k/M} \tag{5.10.45}$$

in which M is the number of distinct filter coefficients, h_k, $k = 1, \ldots, M$ and $\omega_M = (M-1)/M$ is the frequency of the last term kept in the Fourier expansion. Multiplication of the weights of the cosine filter by the sigma factors yields the desired weights of the Lanczos-window cosine filter. Thus, the weights of the low-pass cosine–Lanczos filter become, using $\sigma(M, 0) = 1$

$$h_o = \omega_c/\omega_N, \quad \text{for } k = 0 \tag{5.10.46a}$$

$$h_k = (\omega_c/\omega_N)\frac{\sin(\pi k\omega_c/\omega_N)}{\pi k\omega_c/\omega_N}\sigma(M,k) \tag{5.10.46b}$$

for $k = \pm 1, \ldots, \pm M$ and $M \ll N$. The corresponding weights for the high-pass Lanczos–cosine filter are

$$h_o = 1 - \omega_c/\omega_N, \quad \text{for } k = 0 \tag{5.10.47a}$$

$$h_k = -(\omega_c/\omega_N)\frac{\sin(\pi k\omega_c/\omega_N)}{\pi k\omega_c/\omega_N}\sigma(M,k) \tag{5.10.47b}$$

The transfer function (5.10.39) for a low-pass cosine–Lanczos filter is then

$$H_L(\omega) = \frac{\omega_c}{\omega_N}\left[1 + 2\sum_{k=1}^{M-1}\sigma(M,k)\frac{\sin(\pi k\omega_c/\omega_N)}{\pi k\omega_c/\omega_N}\cos(\pi k\omega/\omega_N)\right] \tag{5.10.48}$$

while for the high-pass cosine–Lanczos filter

$$H_H(\omega) = 1 - H_L(\omega) \tag{5.10.49}$$

Examination of the transfer functions in Figure 5.10.12 reveals that the side-lobe ripples are considerably reduced by the sigma factors of the Lanczos window. Again, the tradeoff is a broadened central lobe, so that, although there is much less contamination from frequencies within the stop-band, the transition of the filter amplitude at the pass-band is less steep than that for the cosine filter. The effect of this smoothing,

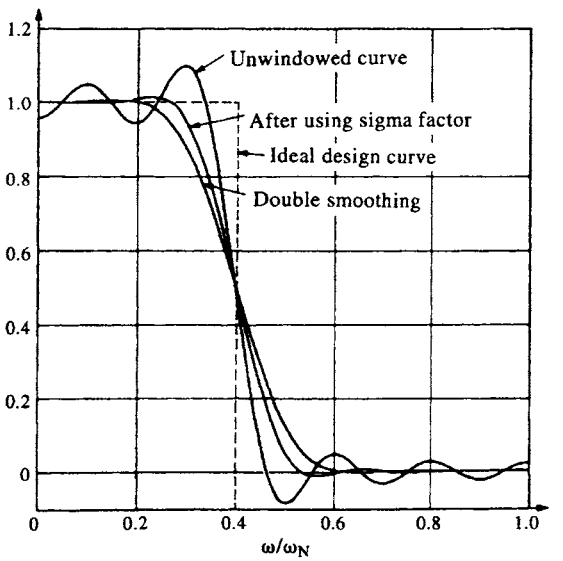


Figure 5.10.12. Approximations to the frequency response of an ideal low-pass filter (dashed line). Solid curves give: The frequency response for an unwindowed cosine filter, a Lanczos–cosine filter that uses sigma factors, and the response after double application of the Lanczos–cosine filter. Filters use $M = 10$ Fourier terms and $\omega_c = 0.4\omega_N$; ω_N = Nyquist frequency. Gibbs' effect is reduced by the sigma factors of the Lanczos window. (From Hamming, 1977.)

which represents a long period modulation of the weighting terms h_k in (5.10.42), can be illustrated numerically by taking a record length $N = 25$ and calculating the filter response $H(\omega/\omega_N)$ with and without the sigma factors. This exercise is instructive in other ways in that it emphasizes the effect of truncation errors during the calculations and indicates what happens if ω_c/ω_N is too near to the ends of the principal interval $0 \leq \omega/\omega_N \leq 1$. Consider the case $\omega_c/\omega_N = 0.022$, $N = 25$, and filter truncation at the fourth decimal place. For a high-pass cosine-type filter with no Lanczos window (which we want to have zero amplitude near zero frequency), we find $H(0) = 0.0740$ whereas use of the sigma factors (Lanczos window) yields $H(0) = 0.4015$. With the cut-off frequency so close to the end of the frequency range, the sigma factors clearly degrade the usefulness of the filter. Increasing the record length to $N = 50$ for the same cut-off frequency improves matters considerably; in this case, $H(0) = 0.0527$ and $H(1) = 0.9997$ using the sigma factors.

5.10.7.3 Practical filter design

Design of a low or high-pass cosine–Lanczos filter begins with specification of: (1) The cut-off frequency; and (2) the number M of weighting terms required to achieve the desired roll-off between the stop and pass-bands. The cut-off frequency is then normalized by the Nyquist frequency, ω_N, obtained from the sampling interval Δt of the time series. As with other types of filters, it is advantageous to keep the normalized cut-off frequency away from the ends of the principal interval

$$0 \leq \omega/\omega_N \leq 1 \tag{5.10.50}$$

The weights h_k are then derived via (5.10.46) and (5.10.47).

Using (5.10.4) and (5.10.8), and assuming an input $\{x_n\}$, $n = 0, 1, \ldots, N-1$, the output for a low-pass cosine–Lanczos filter with $M + 1$ weights is

$$y_n = \frac{2\omega_c}{\omega_N}\left[x_n + \sum_{k=1}^{M} F(k)(x_{n-k} + x_{n+k})\right] \tag{5.10.51a}$$

in which

$$F(k) = \frac{\frac{1}{2}\sin(\pi k/M)}{\pi k/M}\,\frac{\sin(\pi k\omega_c/\omega_N)}{\pi k\omega_c/\omega_N} \tag{5.10.51b}$$

The output time series begins with $y_M = y(M\Delta t)$ corresponding to the first calculable value for the given filter length, M, and the assumption that the input data begin at $x_n = x_o$. That is

$$\begin{aligned} y_M = \frac{2\omega_c}{\omega_N}\Big[&x_M + \frac{1}{2}F(1)(x_{M-1} + x_{M+1}) \\ &+ \frac{1}{2}F(2)(x_{M-2} + x_{M+2}) + \ldots \\ &\ldots + \frac{1}{2}F(M)(x_o + x_{2M})\Big] \end{aligned} \tag{5.10.52}$$

The chosen number of filter coefficients, M, is always a compromise between the desired roll-off of the filter at the cut-off frequency and the acceptable number of data

points (= $2M$) that are lost from the two ends of the record. The greater the number M, the sharper the filter cut-off and the greater the data loss. Repeated (q times) processing of a given record by the same filter generates an increasingly sharper cascade filter response $[H(\omega/\omega_q)]^q$ with an corresponding greater loss (qM) of data values from each end of the record. For a high-pass filter, M should be large enough that, in the time domain, the $2M$ weights for the corresponding low-pass filter span "many" periods of the higher frequency oscillations one is attempting to isolate using the filter.

The sum S of the weights h_k in (5.10.46) and (5.10.47)

$$S = \sum_{k=0}^{M} h_k$$

gives a qualitative measure of the filter performance. An ideal low-pass filter (i.e. one with no truncation or numerical roundoff effects) should give $S = 1$ while an ideal high-pass filter would have $S = 0$. Close proximity to these values indicates a numerically reliable filter.

5.10.7.4 The Hanning (von Hann) window

A variety of cosine-type filters are presented in the recent oceanographic literature under the general term of Lanczos–cosine or cosine–Lanczos filters. A popular formulation having widespread application is the five-day low-pass filter proposed by Mooers and Smith (1967) in a study of continental shelf waves off Oregon. In this study, a Hanning or raised cosine window defined by

$$\begin{aligned} \omega_k &= \tfrac{1}{2}[1 + \cos(\pi k/M)], \quad |k| < M \\ &= 0, \quad |k| > M \end{aligned} \tag{5.10.54}$$

replaces the sigma factors in (5.10.47).

Let x_n, $n = 1, 2, \ldots, N$ denote an hourly digital time series and $2M + 1 = 120$ be the total number of weights spanning a period of 120 h (five days). The hourly output $\{y_n\}$ from the filter is then

$$y_n = \frac{1}{A}\left[x_n + \sum_{k=1}^{60} F(k)(x_{n-k} + x_{n+k})\right] \tag{5.10.55a}$$

where

$$F(k) = \frac{1}{2}[1 + \cos(\pi k/60)]\frac{\sin(p\pi k/12)}{p\pi k/12} \tag{5.10.55b}$$

and

$$A = 1 + 2\sum_{k=1}^{60} F(k) \tag{5.10.55c}$$

is the normalization factor. Once the number of filter weights k is specified (here, $k = 60$), the transfer function $H_L(\omega)$ is determined by the parameter p, the half-amplitude frequency of the filter in cycles per day (cpd). Specifically, we find

$$H_L(\omega) = \frac{1}{A}\left[1 + 2\sum_{k=1}^{60} F(k)\cos(\pi k\omega/\omega_N)\right] \tag{5.10.56}$$

in which F and A are given by (5.10.55b) and (5.10.55c).

Comparison of (5.10.55b) with (5.10.51b) shows that

$$p = 12(\omega_c/\omega_N) = 24f_c \text{ (in cpd)} \tag{5.10.57}$$

where $f_c = \omega_c/2\pi$ is the cut-off frequency in cycles per hour (cph) and where we have used the Nyquist frequency $f_N = 0.5$ cph for the hourly sampled data. The arguments of the angles in (5.10.51) and (5.10.55) are, therefore, identical. Where the filters differ is in the use of the sigma factors. Whereas the oscillations of $[1 + \cos(\pi k/M)]$ are uniform with k, those of $\sin(\pi k/M)/(\pi k/M)$ decay with increasing k, similar to the way we have seen the term, $\sin(\pi\omega_c/\omega_N)/(\pi k\omega_c/\omega_N)$, decay in amplitude (e.g. Figure 5.6.1b). In this regard, the raised cosine window provides a more severe weighting of the truncated Fourier series than the sigma factors.

The value $p = 0.7$ cpd, corresponding to a cut-off period of 34.29 h, has been commonly used in the design of low-pass Lanczos–cosine filters (cf. Bryden, 1979). Although this produces an acceptable filter response for periods of two days and longer (where two days is generally the central period of the oceanic "spectral gap"), it has been shown to pass an unacceptable amount of high-frequency energy from the diurnal band, particularly from the O_1 and Q_1 tidal constituents (Walters and Heston, 1982). In an attempt to further reduce the leakage from the diurnal band, Mooers and Smith (1967) applied a separate filter to the low-pass filtered data from the $p = 0.7$ cpd filter or "Lancz7" filter (Thompson, 1983; Figure 5.10.13). Walters and Heston (1982) passed the data twice through the filter to produce the 10-day (Lancz7) filter. This results in a significantly improved filter amplitude throughout the diurnal band but also doubles the amount of data lost from the ends of the time series. Thompson (1983) suggested the use of a Lanczos–cosine filter with $p = 0.6$ cpd (the "Lancz6" filter) which equates to a cut-off period of 40 h. The Lancz6 filter essentially removes the leakage from the diurnal band but simultaneously shifts the low-pass portion of the filtered record to periods somewhat in excess of two days. The difference in the filters is quite subtle. For the Lanczos–cosine filter with $p = 0.7$ (Lancz7 filter), the first zero of the transfer function occurs at 15.4°/h (at 0.0428 cph), which is past the diurnal band (Figure 5.10.13); for the Lancz6 filter, the first zero is shifted to 14°/h (at 0.0389 cph) near the O_1 frequency of 13.9°/h.

5.10.8 Butterworth filters

The windowed cosine filters described in the previous section attempt to approximate an ideal rectangular transfer function using truncated Fourier cosine series. For nonrecursive filters, the output is a simple linear combination of the data and the role of the window is to attenuate the overshoot ripples created by truncation in the time domain (Gibbs' phenomenon). We now turn to a specific type of recursive filter for which the transfer function is created using a rational function in sines and cosines. Because this is a recursive filter, the output consists of both input data and past values of the output.

Let $w = w(\omega)$ be a monotonically increasing rational function of sines and cosines in the frequency, ω. The monotonic function

$$|H_L(\omega)|^2 = 1/[1 + (w/w_c)^{2q}] \tag{5.10.58}$$

(Figure 5.10.14) generates a particularly useful approximation to the squared gain of an ideal low-pass recursive filter with frequency cut-off ω_c. (Our filter design will eventually require $w(0) = 0$ so that the final version of $H_L(\omega)$ will closely resemble (5.10.58).)

Butterworth filters of the form (5.10.58) have a number of desirable features (Roberts and Roberts, 1978). Unlike the transfer function of a linear nonrecursive filter constructed from a truncated Fourier series, the transfer function of a Butterworth filter is monotonically flat within the pass and stop-bands, and has high tangency at both the origin ($\omega = 0$) and the Nyquist frequency, ω_N. The attenuation rate of $H_L(\omega)$ can be increased by increasing the *filter order*, q. However, too steep a transition from the stop-band to the pass-band can lead to ringing effects in the output due to Gibbs' phenomenon. Since it has a squared response, the Butterworth filter produces zero phase shift and its amplitude is attenuated by a factor of two at the cut-off frequency, for which $w/w_c = 1$ for all q. In contrast to nonrecursive filters, such as the Lanczos–cosine filter discussed in the previous section, there is no loss of output data from the ends of the record; N input values yield N output values. However, we do not expect to get something for nothing. The problem is that ringing distorts the data at the ends of the filtered output. As a consequence, we are forced to ignore output values near the ends of the filtered record, in analogy with the loss of data associated with nonrecursive filters. In effect, the loss is comparable to that from a nonrecursive filter of similar smoothing performance. A subjective decision is usually needed to determine where, at the two ends of the filtered record, the "bad" data end and the "good" data begin.

Butterworth filters fall into the category of physically realizeable recursive filters having the time-domain formulation (5.10.2) with $k = 0, \ldots, M$. They may also be classified as infinite impulse response filters since the effects of a single impulse input can be predicted to an arbitrary time into the future. To see why we expect $w(\omega)$ to be

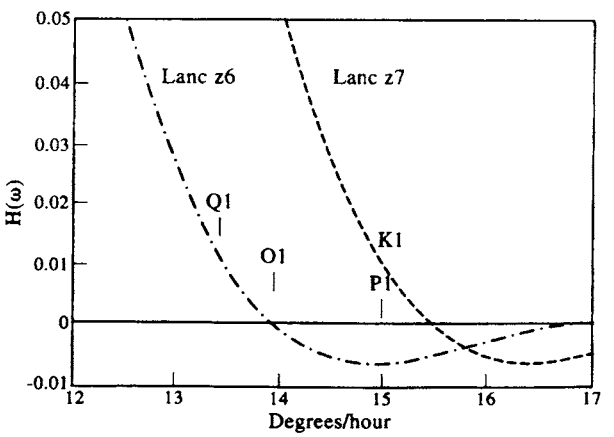


Figure 5.10.13. Expanded views of the filter responses for two tide-elimination filters for the diurnal frequency band. The Lancz6 and Lancz7 filters are low-pass Lanczos–cosine filters. 15°/h = 1.0 cpd. (Modified from Thompson, 1983).

a rational function in sines and cosines, we use (5.10.2) and the fact that $H(\omega)$ is the ratio of the output to the input. We can then write

$$H(\omega) = \frac{\text{output}}{\text{input}} = \frac{\sum_{k=0}^{M} h_k e^{-i\omega k\Delta t}}{1 - \sum_{j=1}^{L} g_j e^{-i\omega k\Delta t}} \tag{5.10.59}$$

where the summations in the numerator and denominator involve polynomials in powers of $\exp(-i\omega k\Delta t)$ which can in turn be expressed through the variable w. The substitution $z = \exp(i\omega kt)$ leads to expression of the filter response $H(\omega)$ in terms of the z-transform and zeros of poles.

5.10.8.1 High-pass and band-pass filters

High-pass and band-pass Butterworth filters can be constructed from the low-pass filter (5.10.58). For example, to construct a high-pass filter with cut-off, w_c, we use the transformation $w/w_c \rightarrow -(w/w_c)^{-1}$ in (5.10.58). The square transfer function of the high-pass filter is then

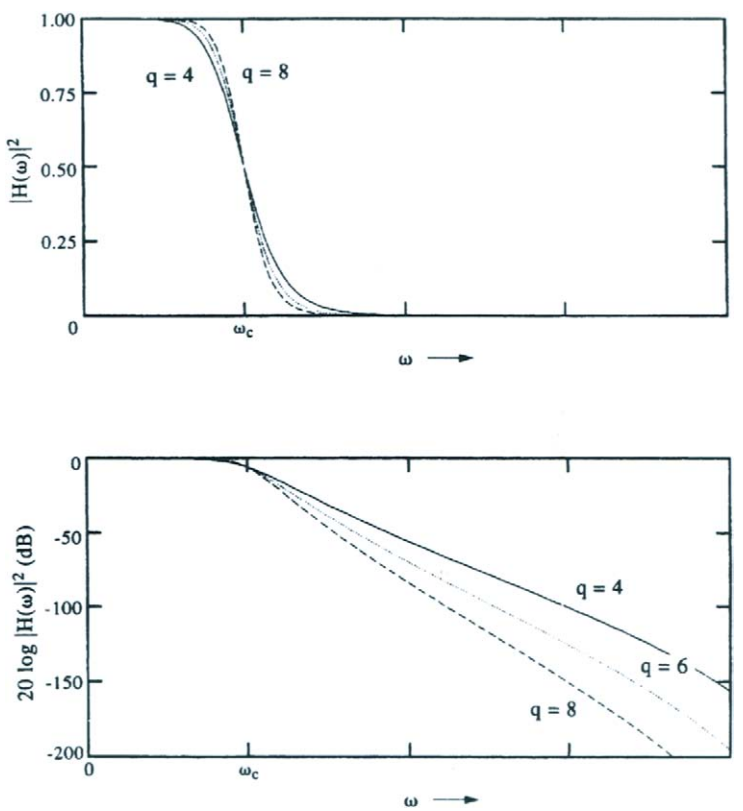


Figure 5.10.14. The frequency response functions $|H_L(\omega)|^2$ for an ideal squared, low-pass Butterworth filter for filter orders q = 4, 6, 8. Bottom panel gives response in decibels (dB). Power = 0.5 at the cut-off frequency, ω_c.

$$|H_H(w)|^2 = (w/w_c)^{2q}/[1 + (w/w_c)^{2q}] \tag{5.10.60}$$

where, as required

$$|H_H(w)|^2 = 1 - |H_L(w)|^2 \tag{5.10.61}$$

Band-pass Butterworth filters (and their counterparts, *stop-band* Butterworth filters) are constructed from a combination of low-pass and high-pass filters. For instance, the appropriate substitution in (5.10.58) for a band-pass filter is $w/w_c = w_*/w_c - (w_*/w_c)^{-1}$ which leads to the quadratic equation

$$(w_*/w_c)^2 - (w/w_c)(w_*/w_c) - 1 = 0 \tag{5.10.62a}$$

with roots

$$w_{*1,2}/w_c = (w/w_c)/2 \pm [(w/w_c)^2/4 + 1]^{1/2} \tag{5.10.62b}$$

Substitution of $w/w_c = \pm 1$ (the cut-off points of the low-pass filter) yields the normalized cut-off functions $w_{*1}/w_c = 0.618$ and $w_{*2}/w_c = 1.618$ of the band-pass filter based on the cut-off frequency $\pm\omega_c$ of the associated low-pass filter. The corresponding band-pass cut-off functions for the cut-off frequency $-\omega_c$ of the low-pass filter are $w_{*1}/w_c = -1.618$ and $w_{*2}/w_c = -0.618$. Specification of the low-pass cut-off determines w_{*1}/w_{*2} of the band-pass filter. The bandwidth $\Delta w/w_c = -(w_{*1} - w_{*2})/w_c = 1$ and the product $(w_{*1}/w_c)(w_{*2}/w_c) = 1$. Note that specification of w_{*1} and w_{*2} gives the associated function w_c of the low-pass filter

$$w_{*1}w_{*2} = w_c^2 \tag{5.10.63}$$

5.10.8.2 Digital formulation

The transfer functions (5.10.58)–(5.10.61) involve the continuous variable w whose structure is determined by sines and cosines of the frequency, ω. To determine a form for $w(\omega)$ applicable to digital data, we seek a rational expression with constant coefficients a to d such that the component $\exp(i\omega\Delta t)$ in (5.10.59) takes the form

$$\exp(i\omega\Delta t) = \frac{aw + b}{cw + d} \tag{5.10.64}$$

(Here, we have replaced $-i\omega\Delta t$ with $+i\omega\Delta t$ without loss of generality.) As discussed by Hamming (1977), the constants are obtained by requiring that $\omega = 0$ corresponds to $w = 0$ and that $\omega \to \pi/\Delta t$ corresponds to $w \to \pm\infty$. Constants b and d (one of which is arbitrary) are set equal to unity. The final "scale" of the transformation is determined by setting $(\omega/2\pi)\Delta t = 1/4$ for $w = 1$. This yields

$$\exp(i\omega\Delta t) = \frac{1 + iw}{1 - iw} \tag{5.10.65}$$

or, equating real and imaginary parts

$$\begin{aligned} w &= \frac{2}{\Delta t}\left[\tan\left(\tfrac{1}{2}\omega\Delta t\right)\right] \\ &= \frac{2}{\Delta t}\left[\tan(\pi\omega/\omega_s)\right], \quad -\omega_N < \omega < \omega_N \end{aligned} \tag{5.10.66}$$

where $\omega_s/(2\pi) = f_s$ is the sampling frequency ($f_s = 1/\Delta t$). We note that the derivation of (5.10.66) is equivalent to the conformal mapping

$$w = \mathrm{i}\frac{2}{\Delta t}\frac{1-z}{1+z} \tag{5.10.67a}$$

where

$$z = \mathrm{e}^{2\pi \mathrm{i} f \Delta t} = \mathrm{e}^{\mathrm{i}\omega\Delta t} \tag{5.10.67b}$$

is the standard z-transform.

The transfer function of the (discrete) low-pass Butterworth filter is then (Rabiner and Gold, 1975)

$$|H_L(\omega)|^2 = \frac{1}{1 + [\tan(\pi\omega/\omega_s)/\tan(\pi\omega_c/\omega_s)]^{2q}} \tag{5.10.68a}$$

and that of the high-pass Butterworth filter

$$|H_H(\omega)|^2 = \frac{[\tan(\pi\omega/\omega_s)/\tan(\pi\omega_c/\omega_s)]^{2q}}{1 + [\tan(\pi\omega/\omega_s)/\tan(\pi\omega_c/\omega_s)]^{2q}} \tag{5.10.68b}$$

The sampling and cut-off frequencies in these expressions are given by $\omega_s = 2\pi/\Delta t$ and $\omega_c = 2\pi/T_c$ in which $T_c = 1/f_c$ is the period of the cyclic cut-off frequency f_c. Plots of (5.10.68a) for various cut-off frequencies and filter order q are presented in Figure 5.10.15.

Use of the bilinear z-transform, $\mathrm{i}(1 - \mathrm{z})/(1 + \mathrm{z})$, in (5.10.67a) eliminates aliasing errors that arise when the standard z-transform is used to derive the transfer function; these errors being large if the digitizing interval is large. Mathematically, the bilinear z-transform maps the inside of the unit circle ($|z| < 1$, for stability) into the upper half plane. A thorough discussion of the derivation of pole and zeros of Butterworth filters is presented in Kanasewich (1975) and Rabiner and Gold (1975).

We note that the above relationships define the square of the response of the filter $H(\omega)$ formed by multiplying the transfer function by its complex conjugate, $H(\omega)^* = H(-\omega)$. (In this instance, $H(\omega)^*$ and $H(-\omega)$ are equivalent since $\mathrm{i} = \sqrt{-1}$) always occurs in conjunction with ω. The product $H(\omega)H(-\omega)$ eliminates any frequency-dependent phase shift caused by the individual filters and produces a squared, and therefore sharper, frequency response than produced $H(\omega)$ alone. The sharpness of the filter (as determined by the parameter q) is limited by filter ringing and stability problems. When q becomes too large, the filter begins to act like a step and Gibbs' phenomenon rapidly ensues.

Equations (5.10.68a, b) are used to design the filter in the frequency domain. In the time domain, we first determine the filter coefficients h_k and g_j for the low-pass filter (5.10.2) and then manipulate the output from the transfer function $H(\omega)$ to generate the output $|H(\omega)|^2$. To obtain the output for a high-pass Butterworth filter, $|H_H(\omega)|^2$,

the output from the corresponding low-pass filter, $|H_L(\omega)|^2$, is first obtained and the resulting data values subtracted from the original input values on a data point-by-data point basis.

5.10.8.3 Tangent versus sine filters

Equations (5.10.68a, b) define the transfer functions of *tangent* Butterworth low-pass filters. Corresponding transfer functions for *sine* Butterworth low-pass filters are given by

$$|H_L(\omega)|^2 = \frac{1}{1 + [\sin(\pi\omega/\omega_s)/\sin(\pi\omega_c/\omega_s)]^{2q}} \tag{5.10.69}$$

where we have simply replaced tan$x$ with sin$x$ in (5.10.68). Although this book deals only with the tangent version of the filter, there are situations where the sine-version may be preferable (Otnes and Enochson, 1972). The tangent filter has "superior" attenuation within the stop-band but at a cost of doubled algebraic computation (the sine version has only recursive terms while the tangent version has both recursive and nonrecursive terms).

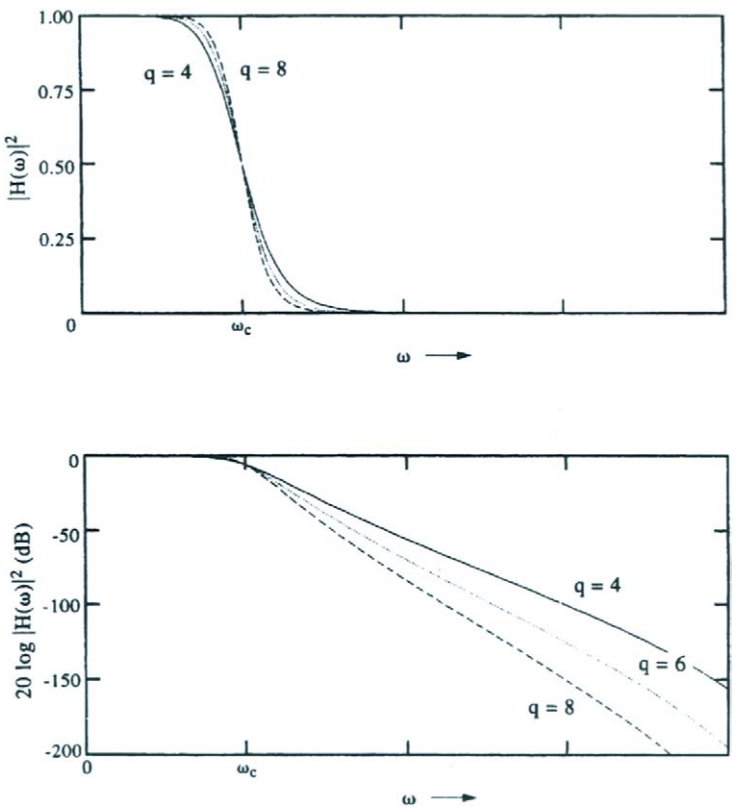


Figure 5.10.15. Same as Figure 5.10.14 but for discrete, low-pass squared Butterworth filters. (After Rabiner and Gold, 1975.)

5.10.8.4 Filter design

The design of Butterworth filters is discussed in Hamming (1977). Our approach is slightly different but uses the same general concepts. We begin by specifying the sampling frequency $\omega_s = 2\pi f_s = 2\pi/\Delta t$ based on the sampling interval Δt for which

$$0 < \omega/\omega_s < 0.5 \tag{5.10.70}$$

and where the upper limit denotes the normalized Nyquist frequency, ω_N/ω_s. We next specify the desired cut-off frequency ω_c at the half-power point of the filter. For best results, the normalized cut-off frequency of the filter, ω_c/ω_s, should be such that the transition band of the filter does not overlap to any significant degree with the ends of the sampling domain (5.10.70). Once the normalized cut-off frequency (or frequencies) is known, specification of the filter order q fully determines the characteristics of the filter response. Our experience suggests that the parameter q should be less than 10 and probably not larger than eight. Despite the use of double precision throughout the calculations, runoff errors and ringing effects can distort the filter response for large q and render the filter impractical.

There are two approaches for Butterworth filter design once the cut-off frequency is specified. The first is to specify q so that the attenuation levels in the pass and stop-bands are automatically determined. The second is to calculate q based on a required attenuation at a given frequency, taking advantage of the fact that we are working with strictly monotonic functions. Suppose we want an attenuation of $-D$ decibels at frequency ω_a in the stop-band of a low-pass filter having a cut-off frequency $\omega_c < \omega_a$. Using the definition for decibels and (5.10.48), we find that

$$\begin{aligned} q &= 0.5 \frac{\log(10^{D/10} - 1)}{\log(w_a/w_c)} \\ &\approx \frac{D/20}{\log(w_a/w_c)}, \quad \text{for } D > 10 \end{aligned} \tag{5.10.71}$$

where D is a positive number measuring the decrease in filter amplitude in decibels (dB) and w is defined by (5.10.66). The nearest integer value can then be taken for the filter order provided that the various parameters (ω_a, D) have been correctly specified and q is less than 10. If the latter is not followed, the imposed constraints are too severe and new parameters need to be specified. The above calculations apply equally to specification of q based on the attenuation $-D$ at frequency $\omega_a < \omega_c$ in the stop-band of a high-pass filter, except that $\log(w_a/w_c)$ in (5.10.71) is replaced by $\log(w_c/w_a)$. Since $\log(x) = -\log(1/x)$, we can simply apply (5.10.71) to the high-pass filter, ignoring the minus sign in front of $\log(1/x)$.

5.10.8.5 Filter coefficients

Once the characteristics of a transfer response have been specified, we need to derive the filter coefficients to be applied to the data in the time domain. We assume that the transfer function $H_L(\omega; q)$ of the low-pass filter can be constructed as a product, or cascade, of second-order ($q = 2$) Butterworth filters $H_L(\omega; 2)$ and, if necessary, one first-order ($q = 1$) Butterworth filter $H_L(\omega; 1)$. For example, suppose we required a filter of order $q = 5$. The transfer function would then be constructed via the cascade

$$H_L(\omega; 5) = H_L(\omega; 1) \times H_{L,1}(\omega; 2) \times H_{L,2}(\omega; 2) \tag{5.10.72}$$

in which the two second-order filters, $H_{L,1}$ and $H_{L,2}$, have different algebraic structure. Use of the cascade technique allows for variable order in the computer code for Butterworth filter programs without the necessity of computing a separate transfer function $H_L(\omega; q)$ each time. This eliminates a considerable amount of algebra and reduces the roundoff error that would arise in the "brute-force calculation" of H_L for each order.

The second-order transfer functions for a specified filter order q are given by

$$H_L(\omega; 2) = \frac{[w_c^2(z^2 + 2z + 1)]}{a_k z^2 + 2z(w_c^2 - 1) + \{1 - 2w_c \sin[\pi(2k + 1)/2q] + w_c^2\}} \tag{5.10.73a}$$

where w and z are defined by (5.10.66) and (5.10.67b)

$$a_k = 1 + 2w_c \sin[\pi(2k + 1)/2q] + w_c^2 \tag{5.10.73b}$$

and k is an integer that takes on values in the range

$$0 \leq k < 0.5(q - 1) \tag{5.10.73c}$$

When q is an odd number, the first-order filter $H_L(\omega; 1)$ must also be used where

$$H_L(\omega; 1) = \left(\frac{w_c}{1 + w_c}\right) \frac{z + 1}{z - \left(\frac{1 - w_c}{1 + w_c}\right)} \tag{5.10.74}$$

Again, suppose that $q = 5$. The transfer function H_L is then composed of the lead filter $H_L(\omega; 1)$ given by (5.10.74) and two second-order filters, for which k takes the values $k = 0$ and 1 in (5.10.73). Note that we have strictly adhered to the inequality in (5.10.73c). The first second-order filter is obtained by setting $k = 0$ in (5.10.73); the second second-order filter is obtained by setting $k = 1$. For $q = 7$, a third second-order for $k = 2$ would be required, and so on.

The next step is to recognize that the first-order function (5.10.74) has the general form

$$H_L(\omega) = \frac{d_o z + d_1}{z - e_1} \tag{5.10.75}$$

and that the second-order function (5.10.73a) has the general form

$$H_L(\omega) = \frac{c_o z^2 + c_1 z + c^2}{z^2 - b_1 z - b_2} \tag{5.10.76}$$

where the sine terms in the coefficients of (5.10.73a) change with filter order q. The coefficients d, e in (5.10.74) are obtained by direct comparison with (5.10.73) while the coefficients b, c in (5.10.76) are obtained through comparison with (5.10.73a).

The recursive digital filters (5.10.2), whose time-domain algorithms have the transfer functions (5.10.75) and (5.10.76) are, respectively

$$y_n = d_o x_n + d_1 x_{n-1} + e_1 y_{n-1} \tag{5.10.77}$$

and

$$y_n = c_0 x_n + c_1 x_{n-1} + c_2 x_{n-2} + b_1 y_{n-1} + b_2 y_{n-2} \tag{5.10.78}$$

Direct comparison of (5.10.75) with (5.10.77) yields the time domain coefficients for the first-order filter; comparison of (5.10.76) with (5.10.78) yields the corresponding coefficients for the second-order filters for each value of k beginning with $k = 0$. In particular, we find, for the first-order filter

$$d_0 = d_1 = \frac{w_c}{1 + w_c}; \quad e_1 = \frac{1 - w_c}{1 + w_c} \tag{5.10.79}$$

and for the second-order filter

$$\begin{aligned} b_1 &= -2w_c/a_k; \quad b_2 = [a_k^{-2}(1 + w_c^2)]/a_k \\ &= c_0 = w_c^2/a_k; \quad c_1 = 2c_0; \quad c_2 = c_0 \end{aligned} \tag{5.10.80}$$

where the coefficients in (5.10.80) change with the parameter k according to the number of second-order filters needed to create the filter of order q.

To apply the $q = 5$ filter, we process the input data x_n ($n = 0, 1, \ldots, N$) by the first-order filter (5.10.77). We then take the output from the first-order filter and process it by the first of the second-order filters (5.10.78) with $k = 0$. The resultant output is then processed by the next second-order filter (5.10.78) with $k = 1$. The sequence y'_n ($n = 0, 1, \ldots$) derived from the three filter applications is the low-pass output for the fifth order Butterworth filter $H_L(\omega; 5)$, as indicated by (5.10.72).

The task is only half complete since our ultimate goal is to remove any filter-induced phase shift by smoothing the data with the squared-response of the filter $|H_L|^2$, given by (5.10.50). The sequence we require is: $\{x_n\}$ yields $\{y'_n\}$ as the output from $H_L(\omega)$; $\{y'_n\}$ yields $\{y_n\}$ as the output from $|H_L(\omega)|^2$. To obtain the output $\{y_n\}$ for the square response of the filter, $|H_L(\omega)|^2$, we need to process the output $\{y'_n\}$, obtained from $H_L(\omega)$, with the filter $H_L(-\omega)$. There are three options: (1) We can separately design $H(-\omega)$, a relatively straightforward task involving some sign changes in (5.10.79) and (5.10.80); (2) we can invert the order of the calculations such that the output $\{y'_n\}$ from $H_L(\omega)$ is passed through the inverted version of this filter. That is, the data from $H_L(\omega)$ are first run through the second-order filter ($k = 1$ for $q = 5$), with the output from this filter passed through the second-order filter ($k = 0$) and finally through the first-order filter; or (3) we can simply invert the chronological order of the data $\{y'_n\}$ and pass the inverted sequence through the original filter $H_L(\omega)$. Since all the data are recorded beforehand, we recommend approach (3). The one caution is that the sequence of the final output must be inverted to regain the original chronological order of the data. In all cases, passing the inverted version of $\{y'_n\}$ through the filter cascade removes any phase shift associated with the first pass which produced $\{y'_n\}$ from $\{y_n\}$. A phase shift $\phi(\omega)$ caused by the first sequence of filters $H_1(\omega) \times H_2(\omega) \times \ldots$ is canceled by the phase shift $-\phi(\omega)$ caused by the second sequence of filters $H_L(-\omega)$.

Computer programs designed to carry out the Butterworth filter operations should assign the output $\{y'_n\}$ from each filter as the new input $\{x_n\}$ to the next filter in the cascade until the output corresponding to the filter $|H_L(\omega)|^2$ is achieved. The last set of output is then chronologically inverted and re-run through the same filter. Following the final set of calculations, the output sequence is inverted to ensure correct ordering in time.

To obtain the results for a *high-pass* Butterworth filter, one further operation is required. The final output $\{y_n\}$ $(n = 0, 1, \ldots)$ from the low-pass filter is subtracted point-for-point from the original input, $\{x_n\}$ $(n = 0, 1, \ldots)$ to create the high-pass filtered data $y_{*n} = x_n - y_n$. The procedure to obtain the low and high-pass Butterworth filters is illustrated schematically in Figure 5.10.16.

5.10.9 Frequency-domain (transform) filtering

The type of digital filtering discussed in the previous sections involves convolution of the time-series data with weighting functions called *impulse response functions* that eliminate selected ranges of frequencies from the data. In the case of Fourier transform filtering, the weights are defined in terms of a Fourier transform window or *frequency response function*, $H(\omega)$, and filtering involves: (1) taking the FFT of the original data set; (2) multiplying the FFT output by the appropriate form of $H(\omega)$ that lets through the frequencies of interest and blocks all the others; and (3) taking the inverse FFT of the result to get back a filtered data set in the time domain. These steps are shown schematically in Figure 5.10.17. As an example, $H(\omega)$ might be a low-pass filter designed to eliminate frequency components with periods $2\pi/\omega$ that are longer than 40 h. Alternatively, $H(\omega)$ could be a "notch" filter used to isolate oscillations centered near the local Coriolis frequency, or a two-notch filter designed to remove energy in the diurnal and semidiurnal tidal bands. Transform methods have been discussed from an oceanographic perspective by Walters and Heston (1982), Evans (1985), and Forbes (1988). As these papers indicate, the choice of an "appropriate" form for $H(\omega)$ is critical to the success of the method. Filtering in the frequency domain is attractive because of its simplicity compared to convolution in the time domain and because it is conceptually more in accord with our objective in filtering; namely, to remove specific periodicities in the data while retaining those of interest. Perhaps contrary to expectation, multiplication of the Fourier transform by a window is not always more computationally efficient than convolution of filter weights with the data (Evans, 1985).

We can outline use of the Fourier transform filtering as follows. Suppose we have a time series $x(t)$ with discrete values $x(n\Delta t) = x_n$, where n is an integer in the range $-N < n \leq N$. The Fourier transform of this time series is

$$X_k = \frac{1}{T} \sum_{n=-N+1}^{N} x_n \exp(-i\omega_k n \Delta t) \tag{5.10.81}$$

where $T = 2N\Delta t$ is the record length and the Fourier frequencies are

$$\omega_k = 2\pi f_k = \frac{2\pi k}{T}, \quad -N < k \leq N. \tag{5.10.82}$$

Let $w(r\Delta t) = w_r$, $-s \leq r \leq s$, represent a set of filter weights whose sum is unity to preserve the series mean and whose distribution is symmetric about $r = 0$ to preserve the phase information in the data. The number of weights, $S = 2s + 1$, is called the span of the filter. Since s points are lost from each end of the input data series, the filtered output series

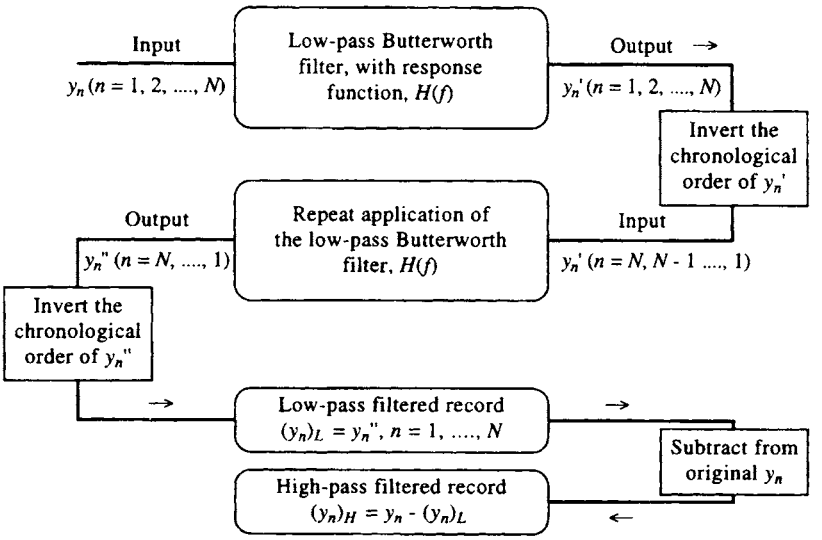


Figure 5.10.16. The procedure for obtaining low and high-pass Butterworth filters.

$$y_n = \sum_{r=-s}^{s} w_r x_{n-r} = \sum_{r=-s}^{s} w_{n-r} x_r \tag{5.10.83}$$

is shorter than the original series by $2s$ values. The effect of the convolution is to smear the signal $x(t)$ according to the weighting imposed by the impulse response function (IRF), $w(t)$. The frequency response function (FRF)

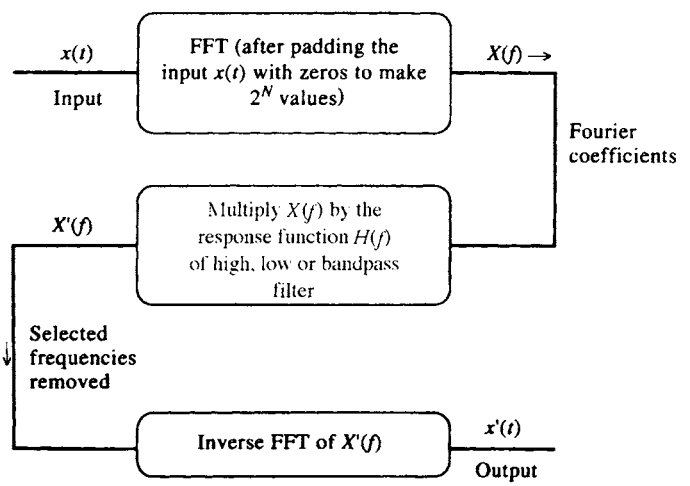


Figure 5.10.17. The procedure for obtaining discrete Fourier transform filters for application in the frequency domain.

$$H(\omega) = \sum_{r=-s}^{s} w_r \exp(-i\omega r \Delta t) = |H(\omega)| e^{-i\phi(\omega)} \tag{5.10.84}$$

gives the effect of the impulse response function on the transform of a sinusoid of unit amplitude and frequency ω ($= 2\pi f$). As stated earlier, the absolute value $|H(\omega)|$ is the *gain factor* of the system and the associated phase angle, $\phi(\omega)$, the *phase factor* of the system. If a linear system is subjected to a sinusoidal input with a frequency ω and produces a sinusoidal output at the same frequency, then $|H(\omega)|$ is the ratio of the output amplitude to the input amplitude and $\phi(\omega)$ is the phase shift between the output and input. The FRF is viewed as a window or transfer function that lets through some frequencies and stops others. Note that H is defined at all frequencies such that $-\pi/\Delta t < \omega \leq \pi/\Delta t$, and not just at the Fourier frequencies, ω_k.

The key to Fourier transform filtering is that, for a constant-parameter linear system, the Fourier transform of the filtered data, $Y(\omega)$, is related to the Fourier transform of the input data, $X(\omega)$, through the product

$$Y(\omega) = H(\omega)X(\omega) \tag{5.10.85}$$

In other words, convolution in the time domain, defined by (5.10.83), translates to multiplication in the frequency domain. The merits of a filter are judged by its FRF (frequency domain) and its IRF (time domain). We would like the magnitude of the FRF to be near unity in the frequency bands to be passed by the filter and near zero in the bands to be stopped; i.e. $|H(\omega)| \approx 1$ and 0, respectively. The transition band between the stop and pass-bands should be as narrow as possible since a broad transition band results in a filtered time series whose frequency content may be contaminated by unwanted frequencies. Similarly, the span of the IRF should be short so that the magnitude of weights decay to zero rapidly as r increases toward $\pm s$. If convolution is used, short filters are computationally more efficient and, moreover, result in less data loss. Unfortunately, the two criteria are at odds with one another. In general, the narrower the transition band in the frequency domain, the slower is the decay rate of the IRF in the time domain. Also, the steeper the maximum slope of the transition band, the larger are the side initial side-lobes of the IRF that arise from the well-known Gibbs' phenomenon. In the limit of a step function-type FRF, in which the transition zone has zero width, the resulting IRF decays very slowly and has large side-lobes (ringing). Thus, one must always compromise in specifying a FRF.

In all time-domain filtering (convolution), data are lost from each end of the original digital time series. For example, in the case of nonrecursive filters, in which the output is based on input time series alone, a known segment of the record of length $T/2$ is lost from either end of the time series (T is the filter length). The same applies to recursive filters in which the present output from the filter is based on the original data series as well as previous values of the output. Here, the difficulty is that the amount of data we must discard from either end is not well defined because of ringing effects associated with the convolution and abrupt data discontinuities at the ends of the record. Transform windowing typically results in exactly the same amount of data loss as the equivalent time-domain filter (Walters and Heston, 1982). The Fourier transform treats the data outside the record as if it were zero, so that the ringing at the ends is introduced by the abrupt changes in the series from nonzero to zero and to the circular convolution of the window's IRF with the data (see Section 5.10.10). Ringing

(Gibbs' phenomenon) occurs throughout the entire time series and becomes evident when the filtered FFT data are inverted to recover the desired filtered time-series data. The effects of Gibbs' phenomenon are mitigated by tapering the frequency-domain filter using a linear or cosine function.

According to Thompson (1983), careful construction of weighting functions in the time domain can more effectively remove tidal components than Fourier transform filtering. This is because tidal frequencies do not generally coincide with Fourier frequencies of the record length. Design of IRF weights to minimize the squared deviation from some specified norm (least squares filter design) offers more control over the FRF at particular nonFourier frequencies. On the other hand, broad-band signals are best served by the FRF approach. Evans (1985) suggests that the ratio of convolution cost to windowing cost is $E = S/[2\log_2(N)]$, where S is the filter span. If $E > 1$, then windowing in the frequency domain is more efficient method. Forbes (1988) addressed the problem of removing tidal signals from the data while retaining the near-inertial signal and argues that Fourier transform filtering is effective provided that careful consideration is given to the filter bandwidth and the amount of tapering of the sides of the filter. Note that, in trying to remove strong tidal signals from a data series, it is sometimes beneficial to first calculate the tidal constituents and then subtract the harmonically predicted tidal signal from the data prior to filtering. This is time consuming and not an advantage if the filter is properly designed.

Figure 5.10.18(a) shows the energy-preserving power spectrum for a mid-depth current meter record from a Cape Howe mooring site (37°35′S, 150°25′E) off the coast of New South Wales. To remove the strong tidal motions from this record, Forbes first used an untapered discrete Fourier transform (DFT) with 12 and 17 adjacent Fourier coefficients set to zero in the diurnal and semidiurnal bands, respectively (Figure 5.10.18b). The greatest improvement in the Fourier transform filtering came from setting only three Fourier terms to zero but tapering the filter with a nine-point cosine taper in the frequency domain at the diurnal and semidiurnal frequencies (Figure 5.10.18c). Thus, tapering the time series, not widening the filter by using more zero frequencies, is a better way to improve filter characteristics. Perhaps, the most important conclusion from Forbes' work is that DFT filters are effective if the number of Fourier coefficients set to zero is sufficient to cover the unwanted frequency band and if the filter is cosine-tapered in the frequency domain to ensure a smooth transition to nonzero Fourier coefficients. In the nonintegral single-frequency case presented here (Forbes was looking at near-inertial motions) this amounted to a three-point filter with a nine-point cosine taper. The widths of the filter and taper must be determined for each application by a careful examination of the spectrum for leakage into adjacent frequencies, but once this is done, the technique is fast and simple to apply.

To summarize the use of Fourier transform filtering:

(1) Remove any linear trend (or nonlinear trend if it is well defined) from the data prior to filtering but do not be too concerned with cosine tapering the first and last 10% of the data. Fast Fourier-transform the data.
(2) Define the Fourier transform filter $H(\omega)$ for both positive and negative frequencies with the extreme frequencies given by $\pm 1/2\Delta t$.
(3) If the measured data are real, and the filtered output is to be real, the filter should obey $H(-\omega) = H(\omega)^*$, where the asterisk denotes complex conjugate. The easiest way to satisfy this condition is to pick $H(\omega)$ real and symmetric in frequency.

(4) If $H(\omega)$ has sharp vertical edges then the impulse response of the filter (the response arising from a short impulse as input) will have damped ringing at frequencies corresponding to these edges. If this occurs, pick a smoother $H(\omega)$. You can take the FFT inverse of $H(\omega)$ to see the impulse response of the filter. The more points used in the smoothing the more rapid the fall of the impulse response.

(5) Multiply the transformed data series $X(\omega)$ by $H(\omega)$ and invert the resultant data series, $Y(\omega)$, to obtain the filtered data in the time domain. To eliminate ringing effects, discard $T/2$ data points from either end of the filtered time series, where T is the span of the IRF for the transform filter.

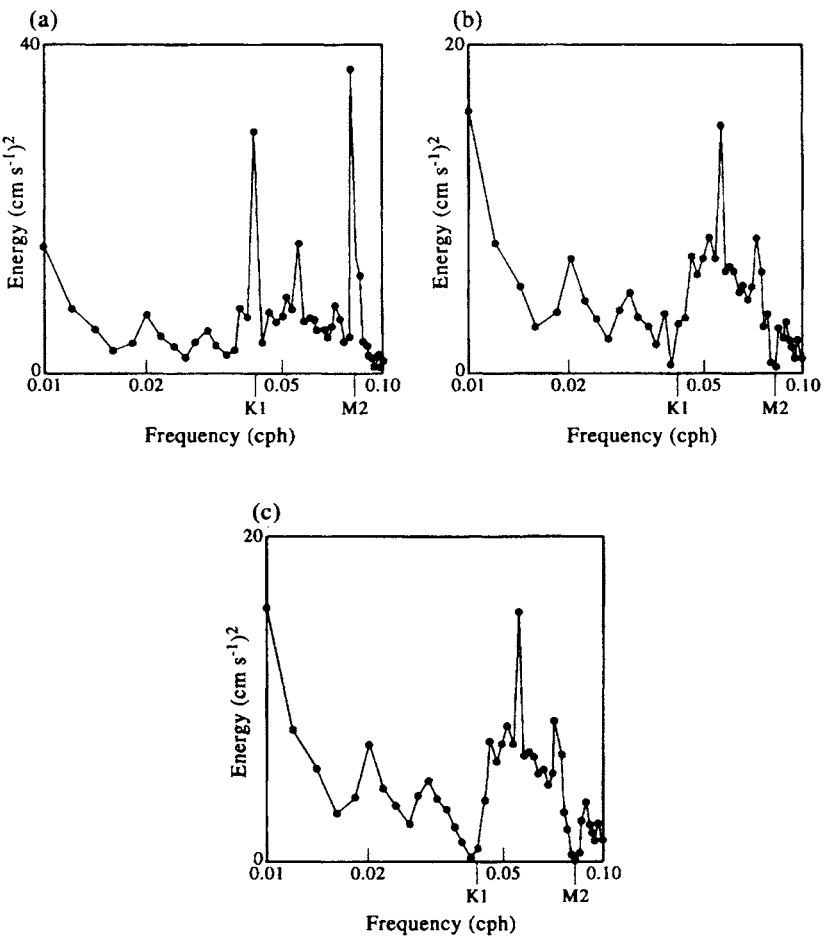


Figure 5.10.18. Energy-preserving spectra for a 4000-h current meter record at 720-m depth off Cape Howe, Australia. (a) Raw hourly data; (b) after applying a discrete Fourier transform (DFT) filter with 12 and 17 adjacent Fourier coefficients set to zero in the diurnal and semidiurnal bands (no tapering); (c) after applying a DFT filter with three Fourier coefficients set to zero and nine Fourier coefficients cosine-tapered on each side of the zero coefficients. (From Forbes, 1988.)

5.10.10 Truncation effects

For all digital filters, a percentage of the end values from the filtered record must be omitted prior to further analysis. This loss of information from the ends of the output is linked to ringing effects associated with discontinuities at the ends of the input and to the nonexistence of integrable data prior to the start of the record. The ringing decays toward the interior of the data sequence after the end effects have been smoothed by a sufficient number of filter integrations (Figure 5.10.19). In the case of the squared Butterworth filter, both ends of the data are affected twice since the data are passed forward and backward through the filter. One approach is to assume that 10% of output data at each end of the filter output is contaminated and remove these points from the final output. However, each case is different and data elimination should be based on a trial and error approach using visual inspection to estimate the extent of the data removal. Padding the ends of the input with zeros appears to serve no useful purpose. In some cases the ringing effect can be substantially reduced by using the zero cross-over points (for input centered about the mean record value) as the first record of the input.

5.11 FRACTALS

The term "fractal" was coined by Mandelbrot (1967) to describe the bumpiness of geometrical curves and surfaces. Regardless of how closely we examine a fractal object it fails to become smooth and its degree of fluctuation remains unchanged. Fractal objects are uneven at all scales and possess no characteristic length scales. Fractals are ubiquitous features whose presence has been reported in a wide variety of fluid dynamical settings including the mixing of turbulent flows (Sreenivasan *et al.*, 1989), the trajectories of oceanic drifters (Osborne *et al.*, 1989; Sanderson *et al.*, 1990) and the paths of atmospheric cyclones (Fraedrich *et al.*, 1990). More everyday examples involve the fractal dimensionality of coastlines, the shapes of clouds, and the forms of lightning strikes. The fractal curve in Figure 5.11.1(a), called a *Koch curve*, resembles a coastline or the outline of a snowflake that would be mapped at ever-increasing spatial resolution. In this case, one begins with an equilateral triangle of side-length L and then successively attaches smaller and smaller equilateral triangles of size $L/3$, $L/3^2$, and so on to the middle of every straight-line segment. After N iterations, the perimeter consists of N segments of length r, where $r = L/3^N$ and

$$N = \alpha(L/r)^D \tag{5.11.1}$$

where $\alpha = 3$ and $D = \log 4/\log 3 \approx 1.262$ is called the fractal dimension. This dimension lies between $D = 1$ for a true one-dimensional curve and $D = 2$ for a true surface area. Figure 5.11.1(b) is an example of an area fractal called the *Sierpinski gasket* which finds use in studies of sediment porosity. Again, one begins with a triangle of side L but then cuts out successively smaller triangles of lengths $L/2$, $L/2^2$, and so on. After N iterations, the "pore" space between the sides of the triangles consists only of triangles of size $r = L/2^N$. The number of such triangles is given by equation (5.11.1) but with $\alpha = 1$ and $D = \log 3/\log 2 \approx 1.585$.

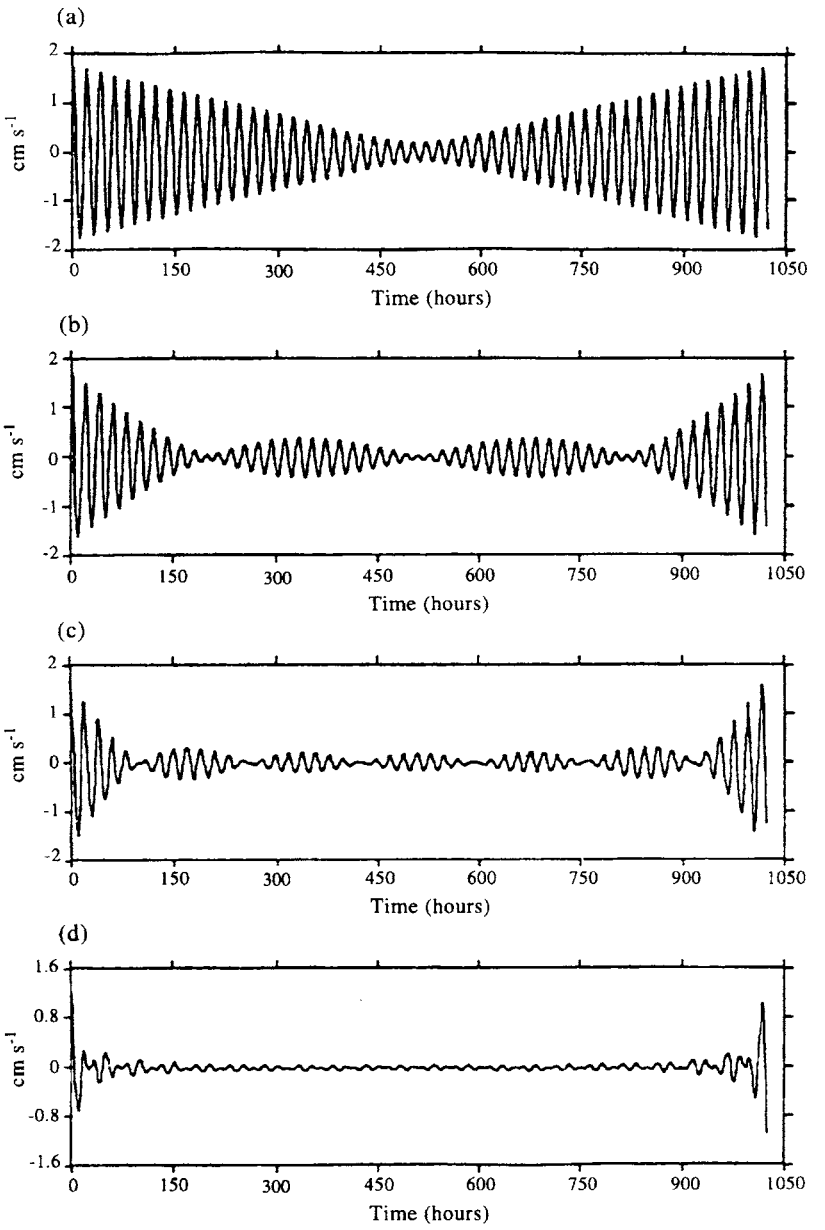


Figure 5.10.19. Ringing effects following application of different discrete Fourier transform filters to an artificial time series with frequency $f = 0.05$ cph and then inverting the transform. (a) Single Fourier coefficient at 0.05 cph set to zero; (b) three Fourier coefficients set to zero; (c) five Fourier coefficients set to zero; (d) 21 coefficients set to zero. (From Forbes, 1988.)

The study of fractal geometry is related to the problem of predictability and propagation of order in nonequilibrium, frictionally dependent dynamical systems, such as turbulent flow in real fluids. In fluid systems, predictability is related to the rate at which initially close fluid particles diverge and the sensitivity of this divergence to initial conditions. Since low predictability implies a highly irregular dynamical system with sensitive dependence on initial conditions, the dispersion of tagged fluid parcels is related to the ultimate skill that can be achieved by deterministic numerical prediction models.

The fractal (or Hausdorff) dimension, D, provides a measure of the roughness of a geometrical object. For example, drifter trajectories confined to a horizontal plane can have a fractal dimension somewhere between that of a topological curve ($D = 1$) and that of random Brownian motion ($D = 2$). The case $D = 1$ is for a smooth differentiable curve whose length remains constant regardless of how the measurements are made. For fractal curves ($D > 1$), the length of the curve increases without bound for decreasing segment length. In the absence of a stationary mean flow, the track of a fluid parcel undergoing Brownian (random-walk) motion will eventually occupy the

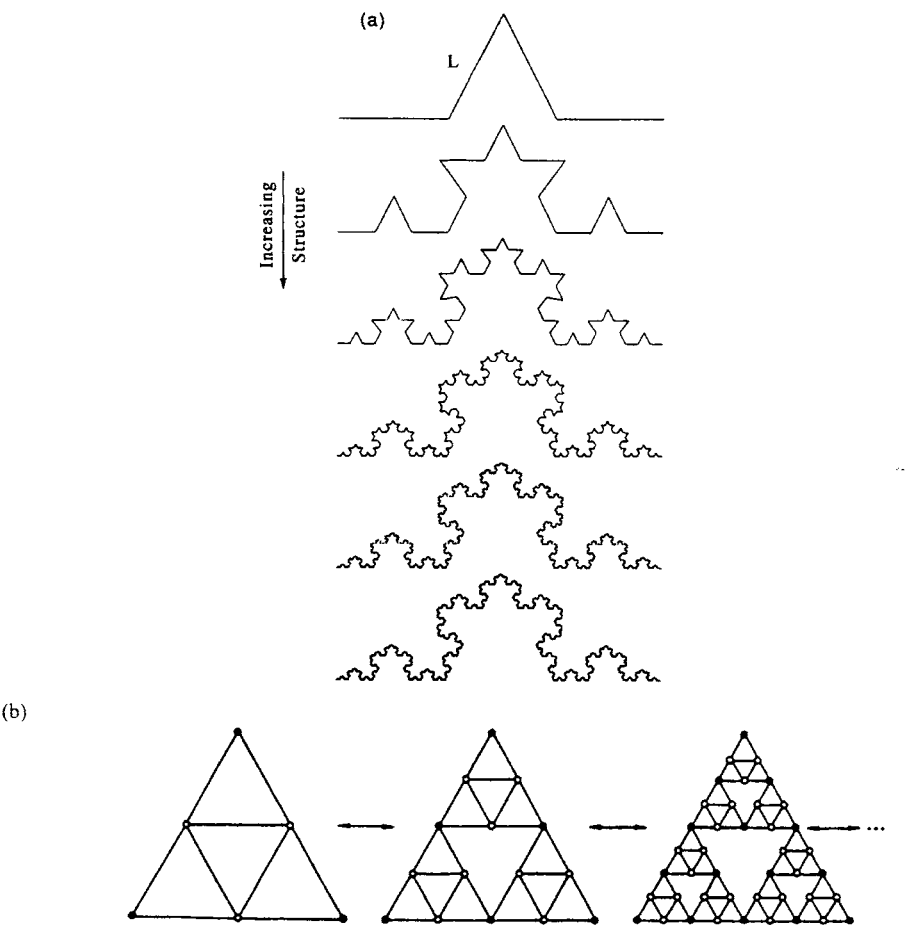


Figure 5.11.1. Examples of common fractals. (a) Generation of the Koch curve fractal by successive attachment of equilateral triangles; D = 1.262; (b) generation of the Sierpinski gasket fractal by successive removal of smaller triangles; D = 1.585.

entire horizontal plane available to it, whereas a parcel displaying fractal Brownian motion will not. The case $D < 2$ implies that the motion has inherent "memory" in the sense that a given incremental displacement in the fluid path is not independent of all previous displacements. In terms of dynamical systems, this means that there are a finite number of variables required to explain the dynamics of the fluid motions.

Osborne *et al.* (1989) examined the scaling properties of drifter trajectories for the upper ocean using year-long tracks of three satellite-tracked drifters deployed within the Kuroshio Extension region in 1977. Based on results from four fundamentally different fractal analysis methods, the Lagrangian trajectories were found to exhibit fractal behavior with dimension $D = 1.27 \pm 0.11$ over spatial scales of 20–150 km and temporal scales of 1.5 days to one week. These scales are thought to be representative of two-dimensional geophysical fluid dynamical turbulence within the inertial subrange—the eddy cascade region of self-similar turbulence which separates short-period current motions (daily tidal oscillations and inertial currents) from long-period oscillations such as Rossby waves and mean flows. Sanderson *et al.* (1990) have reported fractal dimensions at scales of 0.1–4 km for clusters of drifters deployed in Lake Erie, the Atlantic Equatorial Undercurrent, and in coastal waters off the south shore of Long Island. In a related study, the degree of chaotic behavior and predictability of the atmosphere has been studied using tropical and mid-latitude maritime cyclone tracks (Fraedrich and Leslie, 1989; Fraedrich *et al.*, 1990). Results suggest that the atmosphere has an e-folding error growth rate of about 24 h and an ultimate predictability of eight to 14 days.

In this section, we provide several methods for determining the fractal characteristics of oceanic variability using particle track motions.

5.11.1 The scaling exponent method

Consider a particle track sampled at times (t) along the path $\mathbf{x}(t) = (x(t), y(t))$ in longitude–latitude $(x$–$y)$ coordinates. Displacements along each of the two orthogonal horizontal axes are assumed to be independent self-affine (self-scaling) scalar functions. The scaling exponent H (which may be different for the two axes) is positive, less than or equal to unity and related to the fractal dimension of the function by $D = \min[1/H, 2]$. Brownian motions have scaling exponent $H = 1/2$ $(D = 2)$ while monofractal scalar displacements exhibit fractional Brownian motions with $H > 1/2$ $(D < 2)$. If the scalar series are sampled at equal time intervals, the exponents H_x, H_y are given by the *structure functions*

$$\begin{aligned}\overline{[x(t+\alpha\Delta t) - x(t)]^2} &= \overline{[\Delta x(\alpha\Delta t)]^2} \\ &= \alpha^{2H_x}\overline{[\Delta x(\Delta t)]^2} \\ &= \alpha^{2H_x}\overline{[x(t+\Delta t) - x(t)]^2}\end{aligned} \tag{5.11.2a}$$

$$\begin{aligned}\overline{[y(t+\alpha\Delta t) - y(t)]^2} &= \overline{[\Delta y(\alpha\Delta t)]^2} \\ &= \alpha^{2H_y}\overline{[\Delta y(\Delta t)]^2} \\ &= \alpha^{2H_y}\overline{[y(t+\Delta t) - y(t)]^2}\end{aligned} \tag{5.11.2b}$$

where overbars denote averages over time and the α are assigned integer values. The

scaling exponents also can be found using the absolute value of the above functions (Osborne *et al.*, 1989)

$$\overline{|y(t+\alpha\Delta t)-y(t)|} = \alpha^{2H_y}\overline{|y(t+\Delta t)-y(t)|} \tag{5.11.2c}$$

$$\overline{|x(t+\alpha\Delta t)-x(t)|} = \alpha^{2H_x}\overline{|x(t+\Delta t)-x(t)|} \tag{5.11.2d}$$

Figure 5.11.2 provides examples of the scaling exponents, H_y, derived from (5.11.2b) using one-year time series of 6-hourly meridional displacements of 120-m-drogued satellite-tracked drifters launched in the northeast Pacific in 1987. Part (a) of the figure is the log of the structure function

$$\overline{\{|y(t+\alpha\Delta t)-y(t)|^2\}}^{1/2}$$

versus $\log(\alpha)$. The slopes of these curves, H_y, are presented in part (b). Figure 5.11.3 is the same as Figure 5.11.2 except that it uses artificial drifter tracks generated from a Brownian motion (random-walk) algorithm. For the real drifter data, all four tracks had a constant fractal dimension $D_y = 1/H_y \approx 1.18 \pm 0.07$ over time scales of about 0.5–10 days. At longer time scales, motions were strongly affected by mesoscale eddies (cf. Thomson *et al.*, 1990) and fractal analysis is no longer valid. For the pseudo-drifters, $D_y \approx 2$, which is what we would expect for a random-walk regime in which the drifters can occupy the entire two-dimensional space available to them.

Although confined to monofractal functions, the scaling dimension approach is attractive because it is computationally fast and defined in terms of simple scaling properties. The principal drawback is that irregularly sampled particle trajectories, such as those of satellite-tracked drifters, must be converted to equally spaced data using a spline or other interpolation scheme. For isotropic monofractal trajectories, a single fractal dimension is sufficient to define the overall scaling properties of the motions including scaling properties of the mean, variance, and higher moments. Anisotropy in the drifter motions may lead to significantly different values for the scaling exponents H_x, H_y, and associated fractal dimensions. Where these differences are small, fractal dimensions can be expressed through a mean scaling exponent, $\overline{H} = \frac{1}{2}(H_x + H_y)$.

5.11.2 The yardstick method

The fractal dimension of a drifter trajectory of length $L(\Delta)$ can be measured in the usual sense using a ruler (or *yardstick*) with variable length, Δ. As the length of the ruler is decreased and the yardstick estimation of the total length becomes more precise, the length of the trajectory will follow a power-law dependence

$$L(\Delta) \approx \Delta^{1-D_L}; \quad \lim \Delta \to 0 \tag{5.11.3}$$

The divider dimension D_L, which closely approximates the fractal dimension D, is found from the slope of log-transformed $L(\Delta)$ for small length scales Δ (Figure 5.11.4). The case $D_L = 1$ is the topological dimension for a smooth differential curve. For fractal dimensions, $D > 1$ and the length of the curve increases without bound for decreasing segment length.

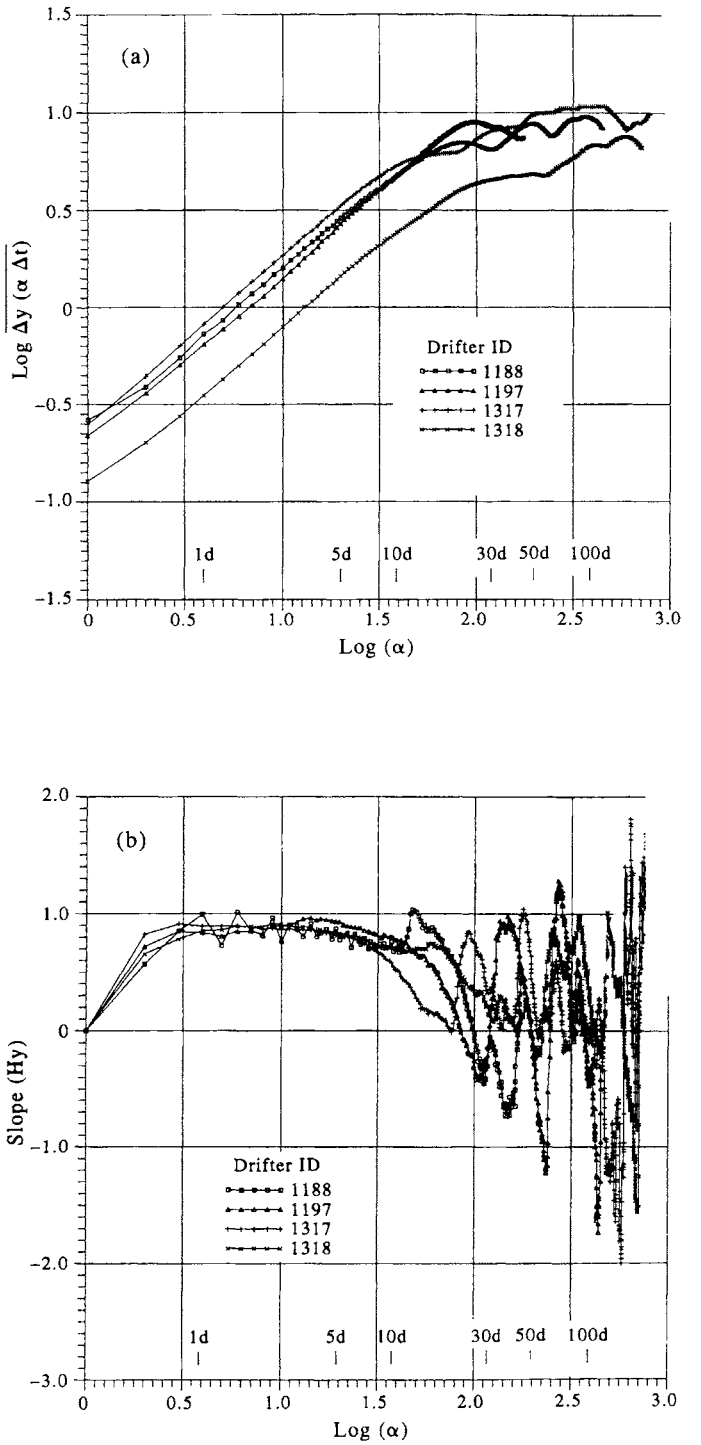
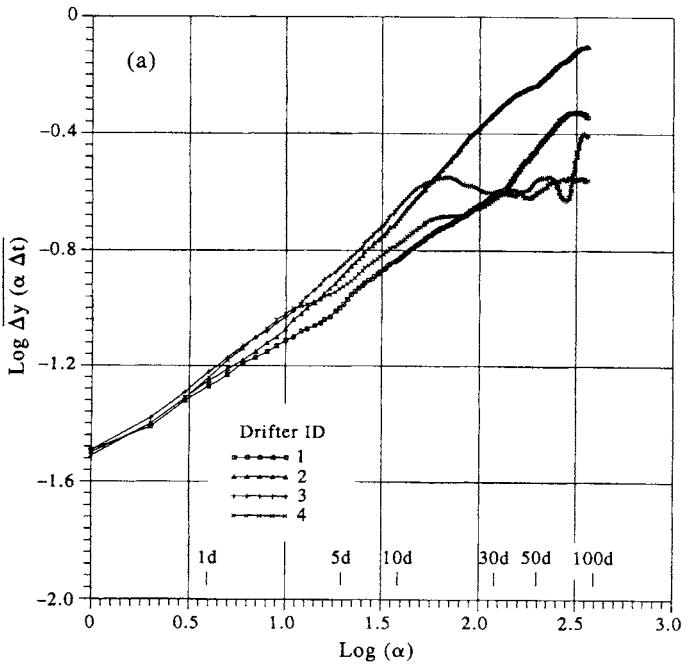


Figure 5.11.2. Structure functions and scaling exponents for trajectories of four 6-hourly sampled, 120-m-drogued satellite-tracked drifters launched in the northeast Pacific in 1987. (a) Absolute values of the structure functions versus the scaling factor, α, plotted on a log–log scale. (b) Slopes, H_y, of the curves in (a) versus scaling factor. Slopes were roughly equal and constant over time scales of one to 10 days.



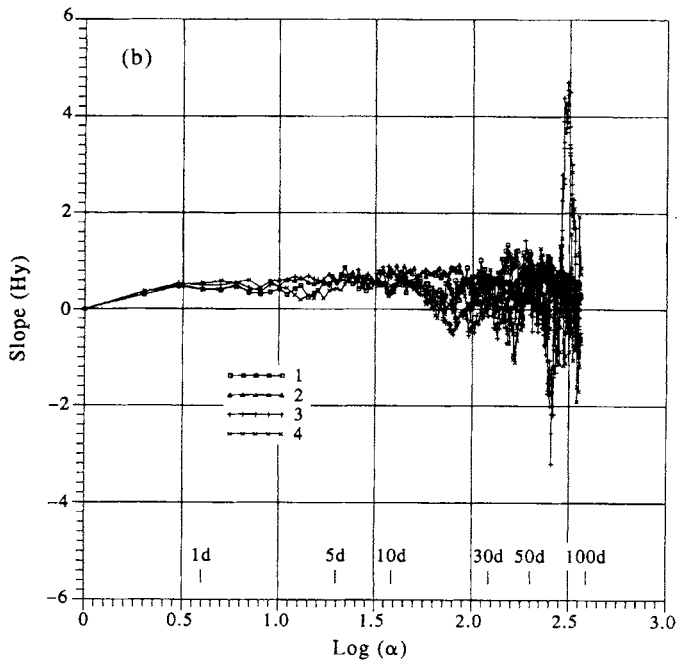


Figure 5.11.3. As in Figure 5.11.2 except for pseudo-drifter tracks generated using a random number generator. In this case, $H_y \approx 0.5$ and drifters perform a non-fractal random walk with dimension $D \approx 2$.

A problem with applying equation (5.11.3) to irregularly sampled drifter records is that the data are unequally sampled both in time and space. Although it makes sense to use a spline-interpolation scheme to generate scalar coordinate data with equally spaced time increments, it is less meaningful to generate coordinate series with equally sampled positional increments. The reason is simple enough: Time is single-valued whereas location is not. Drifters often loop back on themselves. If the data are not equally spaced, we cannot define a sequence of fixed-length yardsticks but must measure the curve $L(\Delta)$ as a function of the average yardstick length, Δ_{av}. This averaging is valid provided the errors introduced by the averaging process are no worse than those arising from other sources (cf. Osborne *et al.*, 1989). Another problem with the yardstick method is that it is based on the slope of (5.11.3) for small spatial scales. The measurement of these scales is often difficult in practice due to limitations in the response and/or positioning of the drifters, cyclone, or other Lagrangian particle.

5.11.3 Box counting method

In this method, one counts the number $N_m(L)$ of boxes of length L in m-dimensional space that are needed to cover a "cloud" or set of points in the space. The Hausdorff–Besicovich dimension, D, of this set can be estimated by determining the number of cubes needed to cover the set in the limit as $L \to 0$. For a fractal curve, the number of boxes increases without bound as $L \to 0$. That is

$$N_m(L) \to L^{-D}, \quad L \to 0 \tag{5.11.4}$$

If the original series is random, then $D = n$ for any dimension n (a random process embedded in an n-dimensional space always fills that space). If, however, the value of D becomes independent of n (i.e. reaches a saturation value, D_0, say), it means that the system represented by the time series has some structure and should possess an attractor whose Hausdorff–Besicovitch dimension is equal to D_0. Once saturation is reached, extra dimensions are not needed to explain the dynamics of the system.

As an example, if we were to measure the area of surfaces embedded in three-dimensional space, we would count the number $N_3(L)$ of cubic boxes of size L required to cover the surface. The area S is then of order

$$S \approx N_3(L)L^2 \tag{5.11.5}$$

For a nonfractal surface, the area asymptotes to a constant value independent of L, which is the true area of the surface. In general

$$N_3(L) \approx L^{-D}, \quad S \approx L^{2-D} \tag{5.11.6}$$

5.11.4 Correlation dimension

An important method for determining the self-similarity of monofractal curves has been proposed by Grassberger and Procaccia (1983). The technique also has found widespread use in studies of chaos and the dimensionality of strange attractors. Specifically, one determines the number of times that the computed distances d_{ij} between points in a time series $x(t_i)$ (or pair of time series $x_i(t)$ and $x_j(t)$) are less than a prescribed length scale, ε. That is, one finds what fraction of the total number of

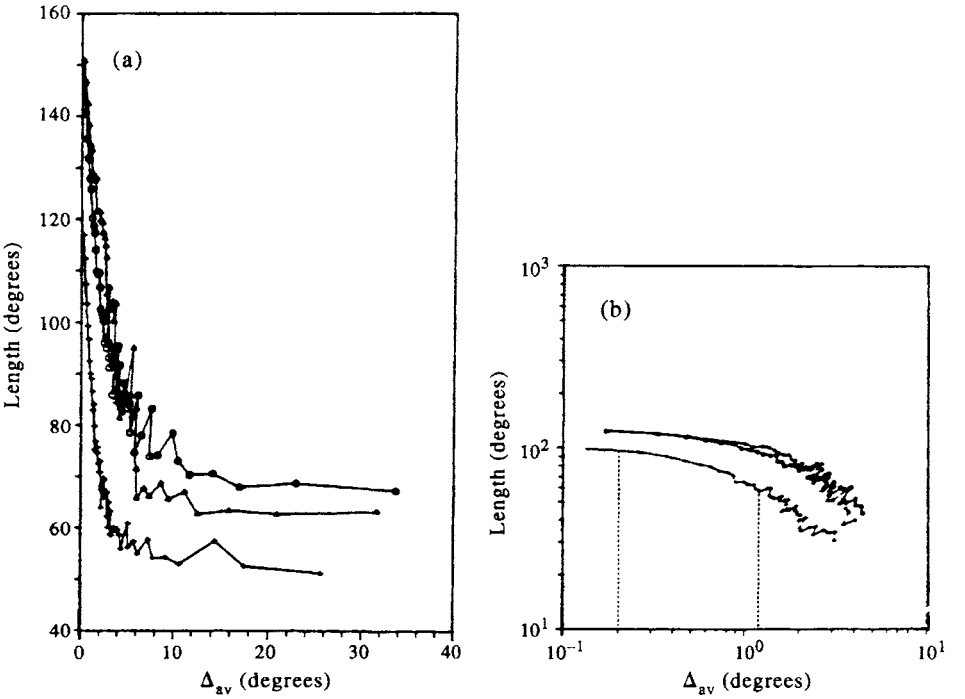


Figure 5.11.4. Yardstick length $L(\Delta)$ measured using a ruler with variable average yardstick length, Δ_{av} (in degrees of latitude), for three drifters launched in the Kuroshio Extension in 1977. (a) Linear coordinates; and (b) log–log coordinates. Note the divergence of the lengths for small Δ. (From Osborne et al., 1989.)

possible estimates of the distance $d_{ij} = |x(t_i) - x(t_j)|$ that are less than ε. For a single discrete vector time series, the Grassberger–Procaccia correlation function is defined as

$$C(\varepsilon) = \frac{1}{M(M-1)} \sum_{ij}^{M} H[\varepsilon - |x(t_i) - x(t_j)|], \quad M \to \infty \tag{5.11.7}$$

where $H(\varepsilon, r_{ij})$ is the Heavyside step function ($= 0$ for $\varepsilon < r$; $= 1$ for $\varepsilon > r$) and M is the number of points in the time series. In (5.11.7), the vertical bars denote the norm of the vector $d_{ij} = [(x(t_i) - x_j)^2 + (y(t_i) - y_j)^2]^{1/2}$. The fractal dimension for a self-affine curve is then obtained as the correlation dimension defined by

$$C(\varepsilon) \approx \varepsilon^{\nu}, \quad \varepsilon \to 0 \tag{5.11.8}$$

The fractal dimension is obtained from the log-transformed version of this equation (Figure 5.11.5). According to Osborne *et al.* (1989), the correlation method gives the least uncertainty in the estimate of the fractal dimension whereas largest errors are associated with the exponent scaling method.

5.11.5 Dimensions of multifractal functions

The various techniques discussed above will (within statistical error) give the same fractal dimension provided that the series being investigated exhibits self-similar monofractal behavior. However, because the techniques rely on different assumptions and measure different scaling properties of the series, the calculated dimensions will be different if the series has a multifractal structure. Multifractal properties are related to multiplicative random processes and are associated with different scaling properties at different scales.

A form of box-counting can be used to study the multifractal properties of ocean drifters (Osborne *et al.*, 1989). Given a fractal curve on a plane, the plane is covered with adjacent square boxes of size Δ and the probability, $p_i(\Delta)$, is computed that the ith box contains a piece of the fractal curve

$$p_i(\Delta) = \frac{n_i(\Delta)}{N} \tag{5.11.9}$$

where n_i is the number of data points falling in the ith box and N is the total number of points in the time series. For fractal curves for small Δ

$$\sum_i [p_i(\Delta)]^q \approx \Delta^{(q-1)D} \tag{5.11.10}$$

where the sum is extended over all nonempty boxes. The quantities $D = D_q$ are the generalized fractal dimensions. A fundamental difference between monofractals and multifractals is that for monofractals D_q is the same for all q while for multifractals the different generalized dimensions are not equal. In general, $D_q < D_{q'}$ for $q > q'$.

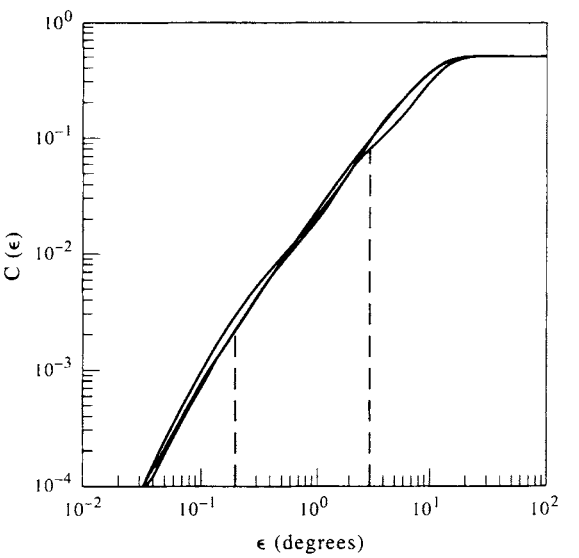


Figure 5.11.5. Correlation functions $C(\varepsilon)$ for three drifters launched in the Kuroshio Extension in 1977. The slope of the function in log–log coordinates is a measure of the correlation dimension of the signal. The two vertical lines indicate the approximate limits of the scaling range. (From Osborne et al., 1989.)

5.11.6 Predictability

A box-counting method can be used to investigate the degree of chaotic behavior associated with the Lagrangian motions such as those of drifters and tropical cyclones. In this method, one counts the number $N_n(\Delta)$ of boxes of dimension Δ in n-dimensional space needed to cover a "cloud" or set of points in the space in the limit $\Delta \to 0$. In practice, the box-counting method is difficult to apply. Estimates of the predictability of drifter trajectories are more readily obtained using the correlation integral technique of Grassberger and Pocaccia (1983). In this case, the degree of predictability is found from the dimension of the attractor derived from an embedded phase space created from all possible pairs of "drifters". The phase space serves, in turn, as a substitute for the state space needed to study the dynamics of a system (Tsonis and Elsner, 1990).

The analysis takes the following steps: (1) we first consider a pair of independent tracks of length $m\Delta t$, where m is the embedding dimension and Δt the sampling increment. Specifically, consider the cyclone tracks for Australia for July 1982 and 1983 (Figure 5.11.6a) examined by Fraedrich *et al.* (1990). For convenience, the start times and positions of the tracks are reinitialized so that they begin at the same time and location. Fraedrich and Leslie (1989) found that the errors introduced by reinitializing are less than those from other sources; (2) we next examine the divergence of the paths by calculating the multiple track correlation function (or correlation integral) $C_m(\varepsilon)$ for the particular embedding dimension m and path separation scale, ε. To this end, we count the number tracks $N_m(\varepsilon)$ of length $m\Delta t$ for which the track length remains less than the great circle distance ε for all the segments in the track. For $m = 1$, each individual data point forms a unit-length segment of the drifter track. One then counts the number of times, $N_1(\varepsilon)$, that the distance between the drifter positions is less than ε for the $N = m$ possible drifter tracks. The distance between each drifter pair is considered; hence, for 10 drifters or cyclone tracks there would be $10 \times 10 = 100$ pairs. This process is repeated for all values of m to create a cloud of points in m-dimensional space which then approximate the dynamics of the system from which the observations $x(t)$ are drawn. The correlation integral is defined by

$$C_m(\varepsilon) = \frac{N_m(\varepsilon)}{[N_m - 1]^2} \tag{5.11.11}$$

where $N_m(\varepsilon)$ is the number of pairs of trajectories of dimension m that remain less than a distance ε from one another. Note that the numerator in the above expression is a squared quantity since it is based on the number of drifter pairs; (3) we then plot $\log [C_m(\varepsilon)]$ versus $ln(\varepsilon)$ to find the slope D_2 of the curve

$$C_m(\varepsilon) \approx \varepsilon^{D_2}, \quad \varepsilon \to 0 \tag{5.11.12}$$

The subscript "2" indicates that pairs of points are used to create the phase space.

If both original time series are random, then $D_2 = 2m$. A random process embedded in a $2m$-dimensional space always fills that space. On the other hand, if D_2 becomes independent of m at some saturation value, D_0, it means that the system represented by the time series has some structure (i.e. predictability) and should possess an attractor whose Hausdorff–Besicovitch dimension is equal to D_0 (Figure 5.11.6b). The need to calculate D_0 from the observations arises because we do not know the value of

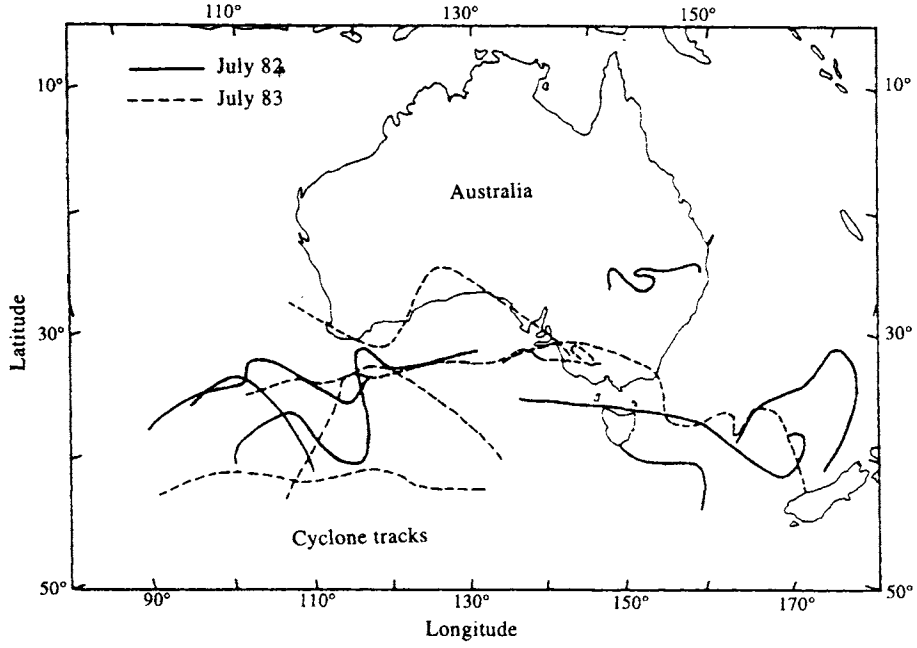


Figure 5.11.6. Use of fractals to study the predictability of cyclone tracks. (a) Cyclone tracks for Australia in July 1982, 1983. (From Fraedrich et al., 1990.)

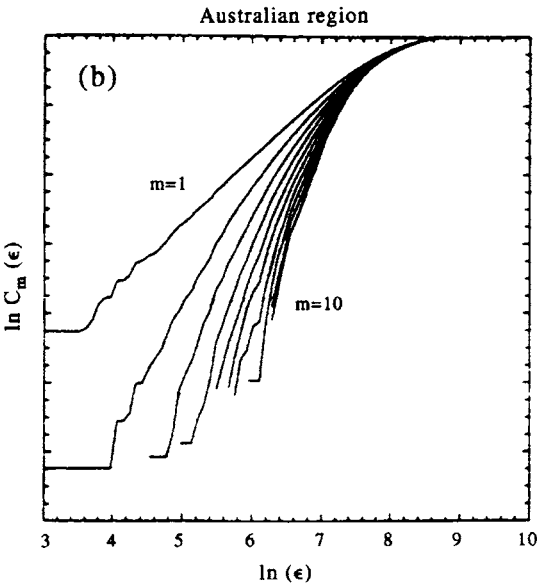


Figure 5.11.6. Use of fractals to study the predictability of cyclone tracks. (b) Correlation integral or cumulative distance distributions $C_m(\varepsilon)$ of pairs of independent cyclone trajectory pieces versus $\ln(\varepsilon)$. Each curve is for m times the data time step of 24 h (m = 1–10 from left to right in the figure). For increasing m, structure eventually becomes invariant at highest embedding dimensions, an indication that extra variables are not needed to account for the dynamics of the system. (From Fraedrich et al., 1990.)

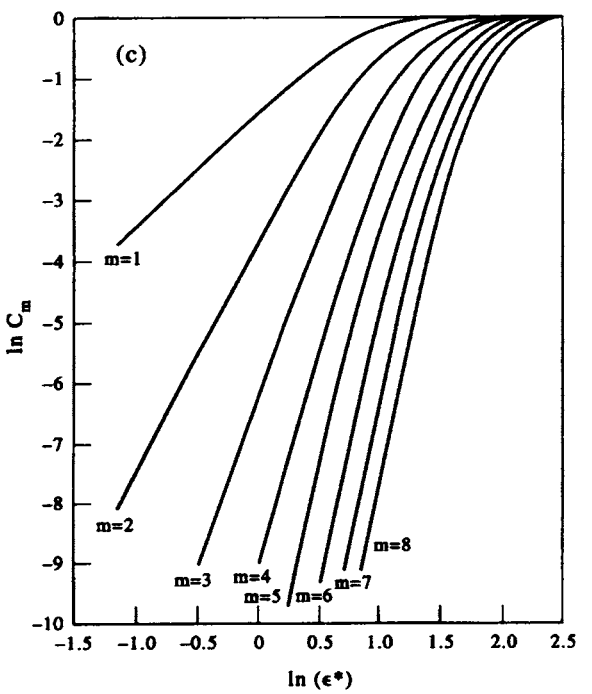


Figure 5.11.6. Use of fractals to study the predictability of cyclone tracks. (c) Same as (b) but for a random-walk pseudo-cyclone generated using a random-number generator. The slopes approach $D_2 \rightarrow 2m$ for decreasing distance threshold, ε. (From Fraedrich et al., 1990.)

m a priori. We, therefore, calculate D_0 for increasing m until we approach a structure that becomes invariant at higher embedding dimensions, an indication that extra variables are not needed to account for the dynamics of the system. The attractor can be a topological structure such as a point, limit cycle or torus, or a nontopological submanifold with fractal structure. For a random-walk regime, D_2 approaches $2m$ so that there is no corresponding limiting value, D_0.

The independent segments of the paired drifter trajectories of sufficiently long duration embed the attractor in a substitute phase space spanned by the time-lagged coordinates provided by the data. The correlation dimension D_2 measures the spatial correlation of the points that lie on the attractor. For a random time series there will be no such spatial correlation in any embedding dimension and thus no saturation will be observed in the exponent D_2. We note that the dimensionality of an attractor, whether fractal or nonfractal, indicates the minimum number of variables present in the evolution of the corresponding dynamical system. In other words, the attractor must be embedded in a state space of at least its dimension. Therefore, the determination of the Hausdorff dimension of an attractor sets a number of constraints that should be satisfied by any numerical or analytical model used to predict the evolution of the system. The main concern is that we do not extend the interpretation when going from a densely populated low-dimensional space to a sparsely occupied high-dimensional space. We cannot go beyond the critical embedding dimension above which the scaling region cannot be accurately determined (Essex *et al.*, 1987; Tsonis and Elsner, 1990).