Time series

In statistics a time series usually means a sequence of observations equally spaced in time.

A time series is wide sense stationary (WSS), if the expectation is independent of the choice of the origin:

$$E(t) = E(t + \tau).$$

A time series is strict sense stationary (SSS), if the probability distribution of the quantity being measured is independent of the choice of the origin.

Analysis of a time series usually consists of the following basic steps:

- Smoothing or other filtering of the observations.
- Removal of the trend or slow variations. The trend can be described e.g. by a straight line or a higher degree polynomial.
- Finding periodic variations. In statistics the variation is often seasonal, or has some other known period.
- Predicting future values and estimating their probability distribution.

Smoothing the data

The simplest method: replace observed values by weighted means

$$g_i = \alpha f_i + (1 - \alpha) g_{i-1},$$

where f_i is the next observed value, g_{i-1} the latest already transformed value and α a constant $0 < \alpha < 1$.

The closer to unity α is, the more closely the resulting curve will follow the original one. When the value is reduced, the curve will become smoother, but at the same time the amplitudi is decreased and there is a phase shift in the variations.

The method works well with real-time applications, since only values that have already been observed are used for smoothing.

Another method is the moving average

$$g_i = \sum_{k=-K}^{K} w_k f_{i+k},$$

where the sum of the weights is $\sum w_k = 1$.

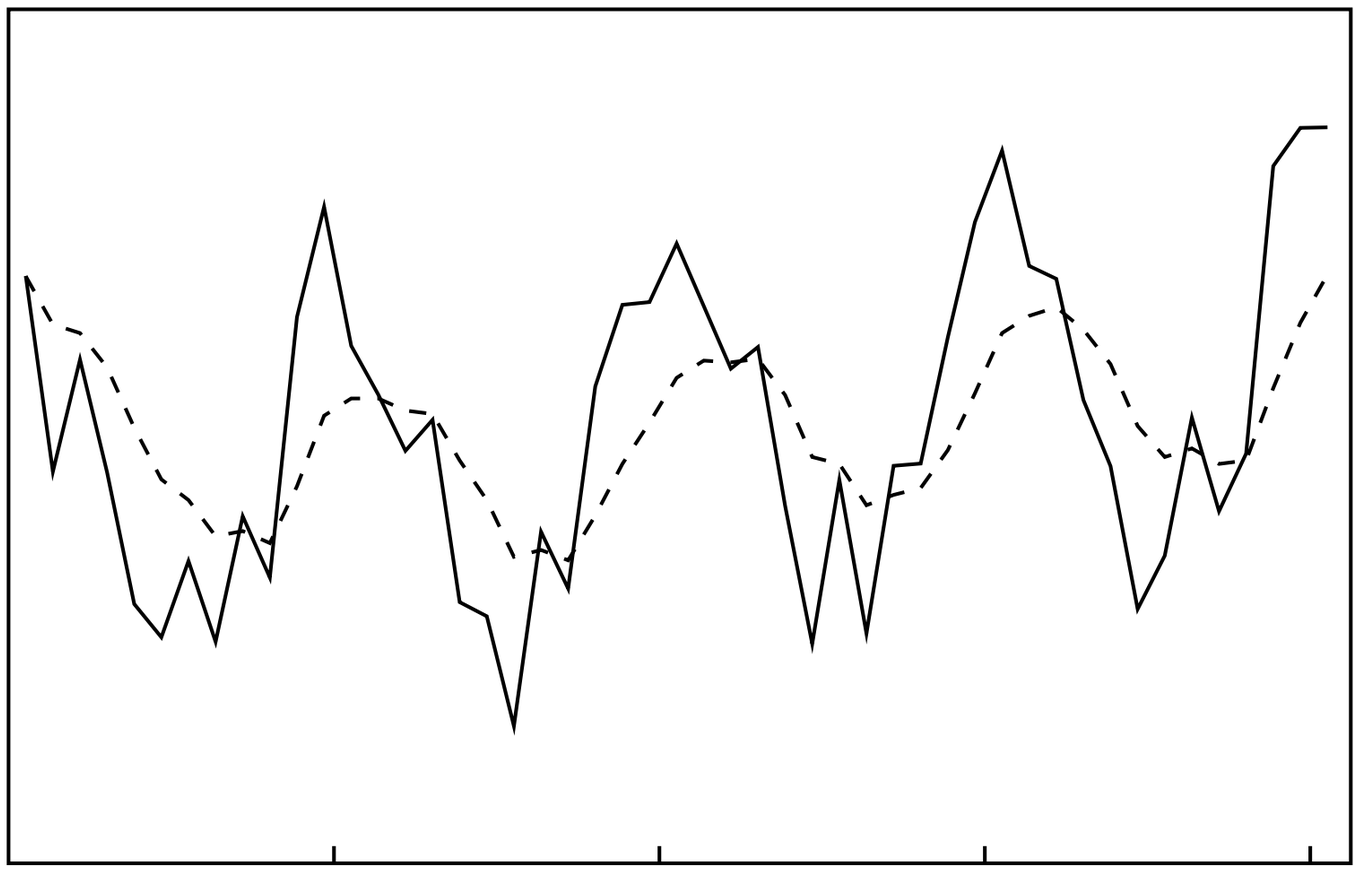
This has no effect on the phase. To obtain as much smoothing as with the previous method, usually more points are needed. The original values cannot be overwritten by the new ones.

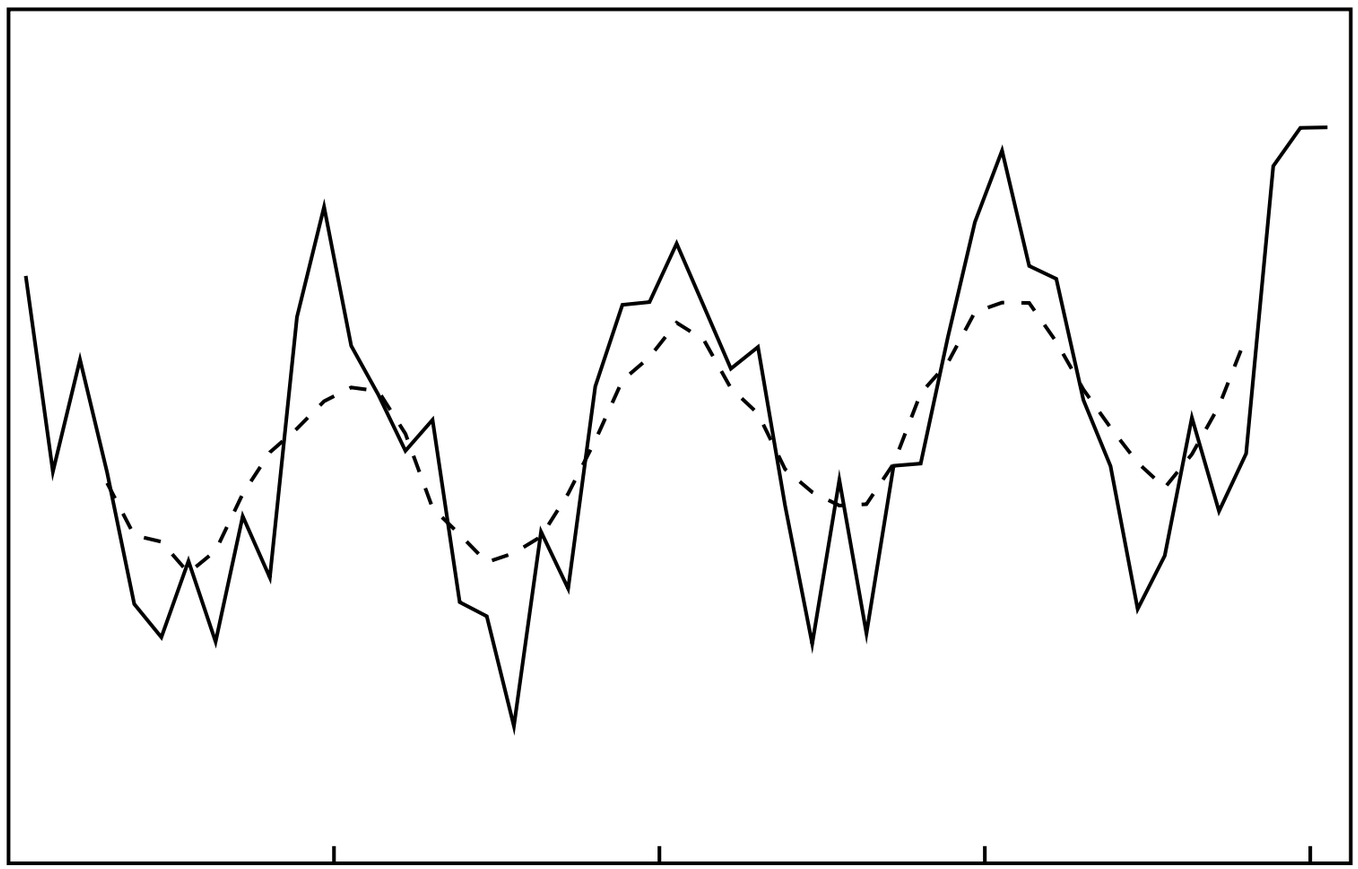
Since both previous and future values are needed, not good for predicting.

The moving average can also be calculated using previous values only

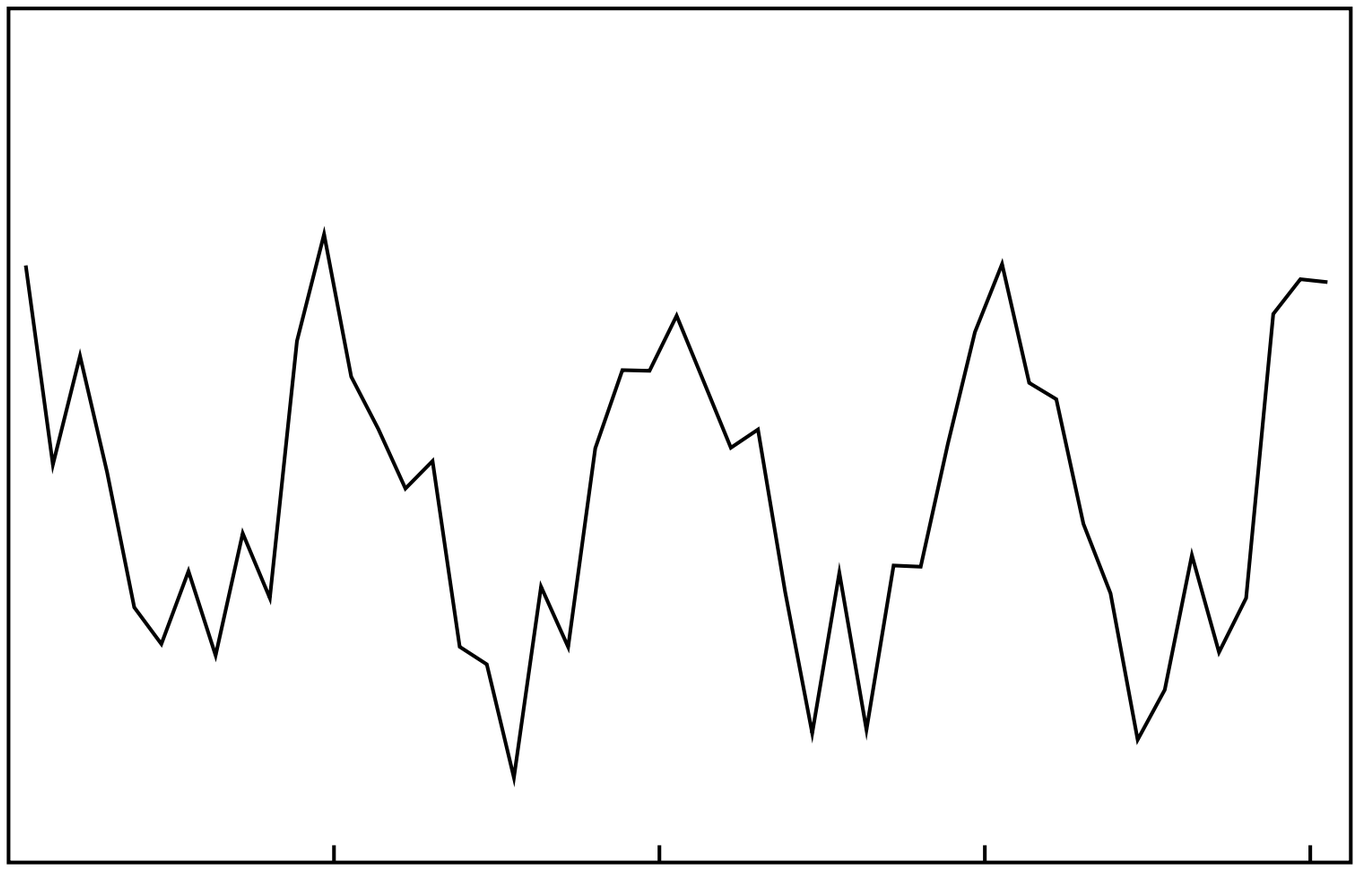
$$g_i = \sum_{k=0}^{2K+1} w_k f_{i-k},$$

which will again cause a phase shift.

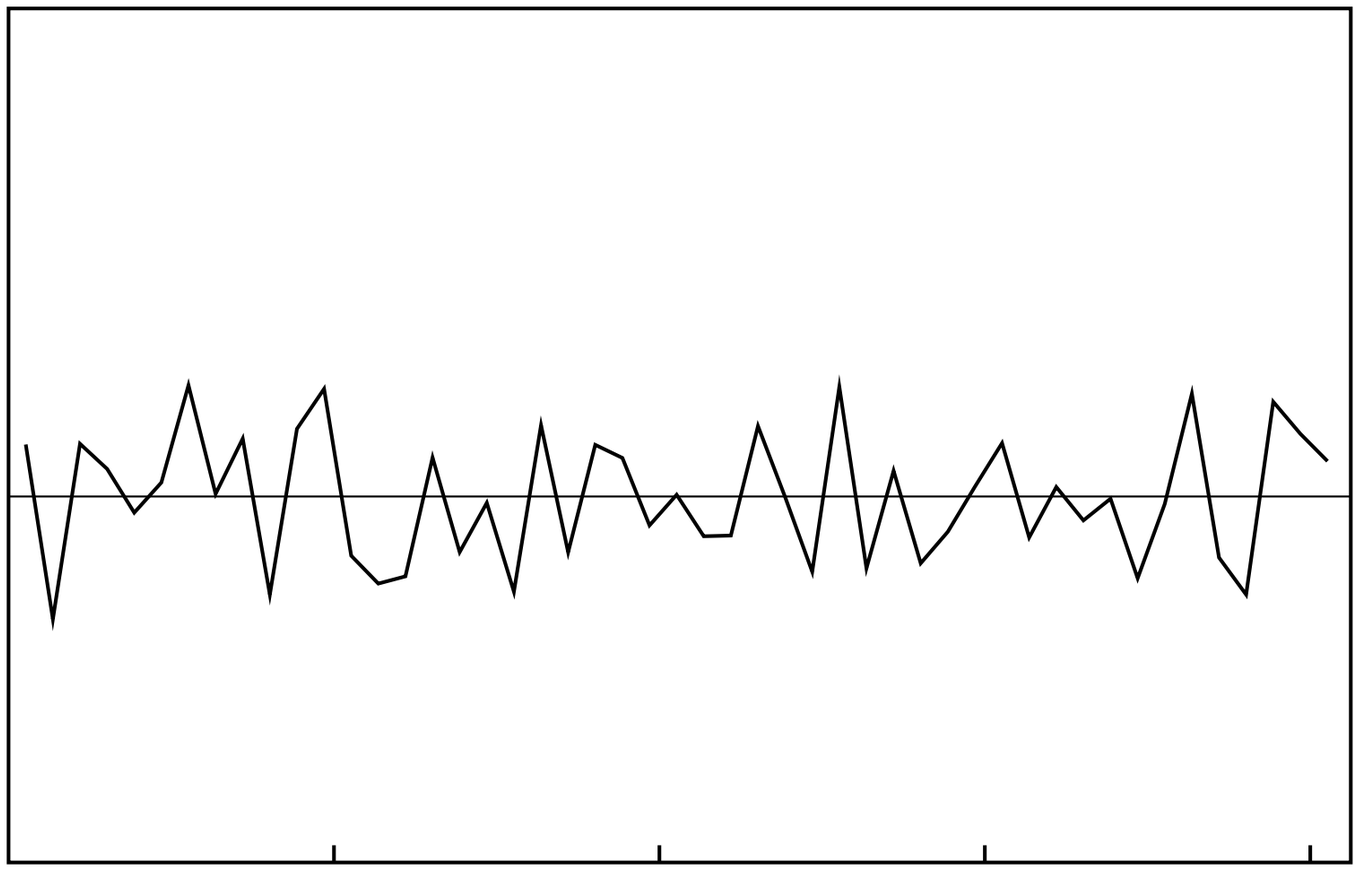




Linear trend removed:



Periodic variation removed:



Regression and autoregression models

Systematic variation can be described by a model depending on time only. The model can also be generalised to a form

$$f(t) = \sum_i a_i \phi_i(t, \mathbf{x}),$$

where the basis functions ϕ_i are arbitrary functions. In addition to time they can depend on any other quantities $\mathbf{x}$, too.

Arguments of the basis functions can be obtained from other time series. The model is then called a regression model.

A regression model can be applied if the phenomena have a real causal relationship that is not just statistical (cf. correlation of icecream sales and drowning accidents).

If a basis function depends on the value of another time series at the same instant of time and we want to evaluate the function, we have to predict the evolution of the other time series, too. (If the shopkeeper of an icecream kiosk wants to order a proper supply for tomorrow, he should predict how many people will get drowned tomorrow.)

If the basis functions depend on the previous values of the function to be predicted, the model is called an on autoregression model, e.g.

$$f(t+1) = 2f(t) - f(t-1),$$

If the data is periodic with a period T, in the simplest case we could use a model

$$f(t+1) = f(t+1-T).$$

If the systematic variation is not sinusoidal, it may be difficult to describe with a few Fourier terms. But a prediction based on an autoregression model can always be found in the same way independently on the form of the variation.

Period determination

Possible problems:

- In addition to the periodicity the data also contains some other variation.
- There are gaps in the data or the time covered by observations is short compared to the actual period.
- The data contains several independent period, which in some cases may lead to erroneous results.
- The period is not constant bu changes with time.
- The phase is changing with time.
- Timing of the observations may cause apparent periods that are not inherent to the phenomenon itself. For example, a variable star is observed a few hours at a time at intervals of one sideral day.

Fourier transform

The Fourier transform gives the frequency components of the data. A maximum indicates periodicity at the coresponding frequency. If the variation is sinusoidal, there is one maximum only. If the variations has a more complex shape, there will be spikes at frequencies that are multiples of the basic frequancy.

If the data contains two or more nearly similar frequancies, the results may be erroneous.

Autocorrelaaion

Dependence between the current observation and the value observed k time steps later is described by the autocorrelation

$$R(k) = \frac{1}{N-k-1} \sum_{i=1}^{N-k} f_i f_{i+k},$$

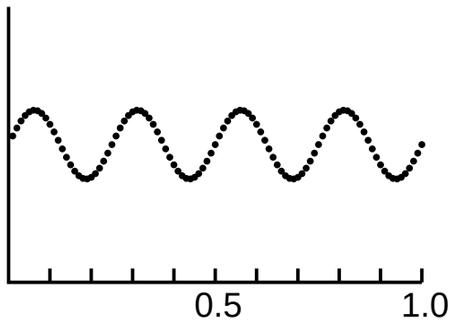
where N is the number of observations. When this is calculated for different values of the delay k, $k = 1, \ldots, N-2$, we get the autocorrelation function. This is a meaningful concept only if the time series is at least wide sense stationary.

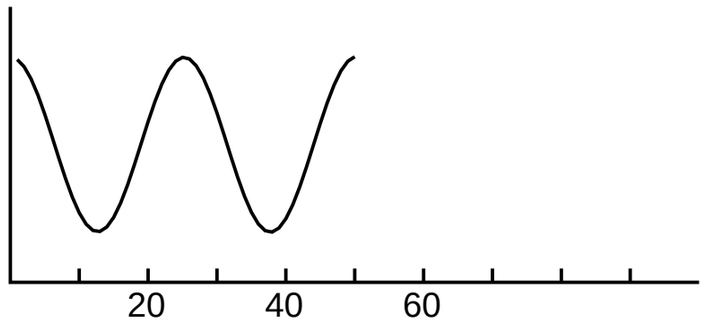
Before evaluating the autocorrelation the trend should be removed from the data.

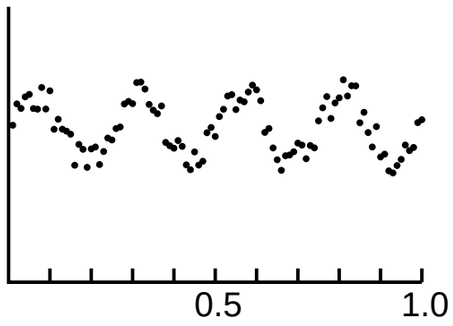
Multiples of the period will show up in the autocorrelation function.

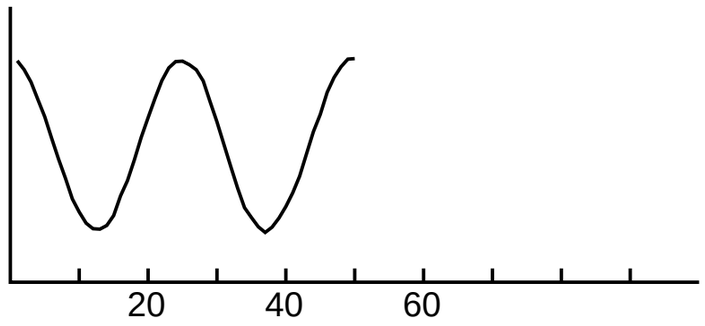
If the observed values consist of pure white noise, successive values are not correlated at all. If the data varies with time, successive values are close to each others, and the strongest correlation is found between successive ($k = 1$) points. Strong correlations corresponding to small values of k are not due to any periodicity. It is only the next local maximum that is an indication of a periodic phenomenon.

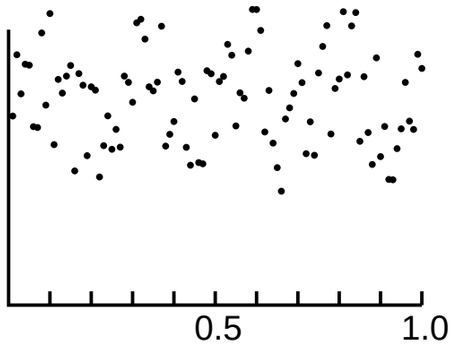
$x = \sin 8\pi t$

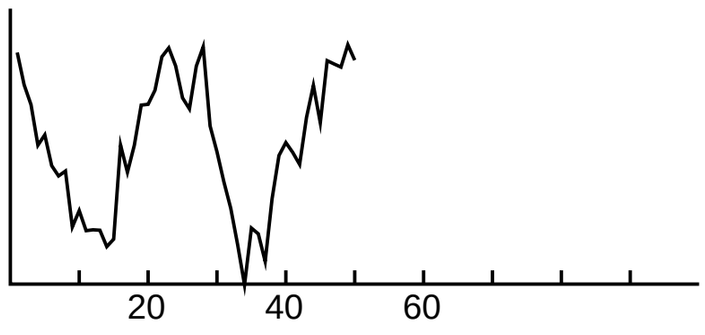












Structure functions

Autocorrelation is a measure for the dependence between the values $x(t+\tau)$ and $x(t)$ in terms of the expectation of their products. Instead of products, also the difference $x(t+\tau) - x(t)$ can be used.

The first degree structure function is

$$D(\tau) = \frac{1}{N} \sum \left(x(t+\tau) - x(t) \right)^2 .$$

tai

$$D(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} \left(x(t+\tau) - x(t) \right)^2 dt.$$

Higher degree structure functions:

$$D_n(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} \left(\sum_{k=1}^{n} (-1)^k \binom{n}{k} x(t+k\tau) \right)^2 dt.$$

The second degree structure function is known as the Allan variance, used widely to describe the stability of the system.

It suffices to have stationary differences. The structure function van be determined even if the autocorerlation function as a function of the delay is not meaningful.

The structure function can be determined even if the mean value and variance are not well defined. Thus it can be applied more generally than the autocorrelation function.

Phasedispersion method

Compare values at a distance T from each others. If they are similar, a small value is obtained for the dispersion measure

$$d(T) = \frac{1}{S} \sum_i \sum_j w(t_i, t_j) |y_i - y_j|^2,$$

where

$$S = \sum_i \sum_j w(t_i, t_j)$$

The function w is chosen in such a way that it is approximately zero except if the time difference is close to the period or its multiple, i.e. $w(t_i, t_j)$ is clearly positive only when $|t_i - t_j| \approx kT$, $k = 0, 1, 2, \ldots$. When the dispersion d is plotted as a function of the period T, we get a curve, whose minima correspond to the periods of the time series.

Unequally spaced data

Astronomical observations are not usually equally spaced.

If there are only a few missing points and the variations are small, the missing points can be interpolated. Generally, this is not usually reasonable.

If the data contains long gaps, it may be necessary to analyze short and long timescales separately using methods designed for equally spaced data. Density of observations can to some extent be smoothed by dividing the observations into bins of suitable length and replacing the data by mean values in each bin.

Autocorrelation function: (Edelson and Krolik, Ap. J., 333, 646, 1988). For each observed time difference there is only one measurement, and the autocorrelation function cannot be properly estimated. A possible solution is to form finite time classes.

Fourier transform; a classical article is: Deeming: Fourier analysis with unequally-spaced data, *Astrophysics and Space Science* **36** 137–158 (1975).

Power spectrum: Scargle, Ap.J., 263, 835, 1982; Lomb, Ap. Space Sci., 39, L147, 1976.

$$P(\omega) = \frac{1}{2}\left(\frac{\left[\sum x_i \cos(\omega(t_i - t_0))\right]^2}{\sum \cos^2(\omega(t_i - t_0))} + \frac{\left[\sum x_i \sin(\omega(t_i - t_0))\right]^2}{\sum \sin^2(\omega(t_i - t_0))}\right),$$

where

$$t_0 = \frac{1}{2\omega} \arctan\left(\frac{\sum \sin 2\omega t_i}{\sum \cos 2\omega t_i}\right).$$

```
program lomb
 implicit none
 integer, parameter :: n=100
 real, parameter :: pi=3.141592654
 real, dimension(n) :: t, x, noise, f, P
 real :: w, k=4.0
 integer i
 read (*,*) w
 call random_number(t)
 call random_number(noise)
 x=sin(k*2*pi*t)+w*noise
 do i=1,n
  f(i)=0.1*i
  P(i)=period(t, x, n, f(i))
  write(*,*) f(i),P(i)
 end do
contains

real function period(t, x, n, f)
! Lombin periodogrammi taajuudella f
! t(1:n)=aika, x(1:n)=havaitut arvot
 integer, intent(in) :: n
```

```
real, dimension(n), intent(in) :: t, x
real, intent(in) :: f
real, parameter :: pi=3.141592654
real s1, s2, s3, s4, mean, var, tau, omega
integer i

omega=2*pi*f
! x:n keskiarvo ja varianssi
mean = sum(x)/n
var= sum((x-mean)**2)/(n-1)
! tau
tau = sum(sin(2*omega*t))/sum(cos(2*omega*t))
tau = atan(tau)/(2*omega)
! P(omega)
 s1 = sum((x-mean)*cos(omega*(t-tau)))
 s2 = sum(cos(omega*(t-tau))**2)
 s3 = sum((x-mean)*sin(omega*(t-tau)))
 s4 = sum(sin(omega*(t-tau))**2)
 period = (s1**2/s2 + s3**2/s4)/(2*var)
end function
end program
```

$x = \sin 8\pi t$