Technical Assistance
HYDROLOGY PROJECT
World Bank & Government of The Netherlands funded

Training module # SWDP - 44

How to carry out correlation and spectral analysis

New Delhi, February 2002

CSMRS Building, 4th Floor, Olof Palme Marg, Hauz Khas,
New Delhi – 11 00 16 India
Tel: 68 61 681 / 84 Fax: (+ 91 11) 68 61 685
E-Mail: hydrologyproject@vsnl.com

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

Table of contents

1. Module context

While designing a training course, the relationship between this module and the others, would be maintained by keeping them close together in the syllabus and place them in a logical sequence. The actual selection of the topics and the depth of training would, of course, depend on the training needs of the participants, i.e. their knowledge level and skills performance upon the start of the course.

2. Module profile

Title	:	Correlation and Spectral Analysis
Target group	:	HIS function(s): ……
Duration	:	x session of y min
Objectives	:	After the training the participants will be able to*:*
Key concepts	:	•
Training methods	:	Lecture, exercises
Training tools required	:	Board, flipchart
Handouts	:	As provided in this module
Further reading and references	:	

3. Session plan

No	Activities	Time	Tools
1	***Preparations***		
2	***Introduction***:	min	OHS x
	Exercise	min	
	Wrap up	min	

4. Overhead/flipchart master

5. Handout

Add copy of the main text in chapter 7, for all participants

6. Additional handout

These handouts are distributed during delivery and contain test questions, answers to questions, special worksheets, optional information, and other matters you would not like to be seen in the regular handouts.

It is a good practice to pre-punch these additional handouts, so the participants can easily insert them in the main handout folder.

7. Main text

Contents

Correlation and Spectral Analysis

1 Introduction

In this module three tools will be discussed which are used to investigate the correlation structure of time series and to identify the major harmonic components available in the time series:

1. Autocovariance and autocorrelation function
2. Crosscovariance and crosscorrelation function, and
3. Variance spectrum and spectral density function.

The estimation procedures are presented with applications.

2 Autocovariance and Autocorrelation function

The autocovariance function of a series Y(t) is defined as:

$$\gamma_{YY}(k) = \mathrm{Cov}[Y(t), Y(t+k)\} = E[(Y(t) - \mu_Y)(Y(t+k) - \mu_Y)] \quad (1)$$

where: $\gamma_{YY}(k)$ = covariance at lag k
μ_Y = mean of series Y

The covariance at lag 0 is the variance of Y:

$$\gamma_{YY}(0) = \sigma_Y^2 \quad (2)$$

To make covariance functions of different processes with different variances comparable the autocovariance is scaled by the covariance at lag 0, i.e. variance. The resulting function is called the autocorrelation function $\rho_{YY}(k)$ and its graphical presentation is generally known as the autocorrelogram:

$$\rho_{YY}(k) = \gamma_{YY}(k))/ \gamma_{YY}(0) \qquad \text{with } -1\leq \rho_{YY}(k) \leq 1 \quad (3)$$

For a random series, i.e. when no serial correlation is present: $\rho_{YY}(k) = 0$ for $|k| > 0$.

In HYMOS the autocovariance function $\gamma_{YY}(k)$ is estimated by the following estimator $C_{YY}(k)$:

$$C_{YY}(k) = \frac{1}{n}\sum_{i=1}^{n-k}(y_i - m_y)(y_{i+k} - m_y) \quad (4)$$

and the autocorrelation function $\rho_{YY}(k)$ by $r_{YY}(k)$:

$$r_{YY}(k) = \frac{C_{YY}(k)}{C_{YY}(0)} = \frac{\frac{1}{n}\sum_{i=1}^{n-k}(y_i - m_y)(y_{i+k} - m_y)}{\frac{1}{n}\sum_{i=1}^{n}(y_i - m_y)^2} \quad (5)$$

Note that the estimators (4) and (5) are biased estimators, since the sum is divided by n rather than by (n-k-1) to obtain estimates with a smaller mean square error.

The 95% confidence limits for zero correlation are according to Siddiqui (see e.g. Yevjevich, 1972):

$$CL_{\pm}(k) = -\frac{1}{n-k+1} \pm 1.96 \frac{(n-k-1)}{(n-k+1)} \sqrt{\frac{1}{n-k}} \qquad (6)$$

Note that the confidence limits are not symmetrical about zero. The confidence region expands slightly with increasing lag k. Now, if the serial correlation for a particular lag k falls within the confidence limits, then for that lag the correlation is not significant.

Example 1

In Example 1 the autocovariance and autocorrelation function is computed for monthly rainfall series of station PATAS, shown in Figure 1.

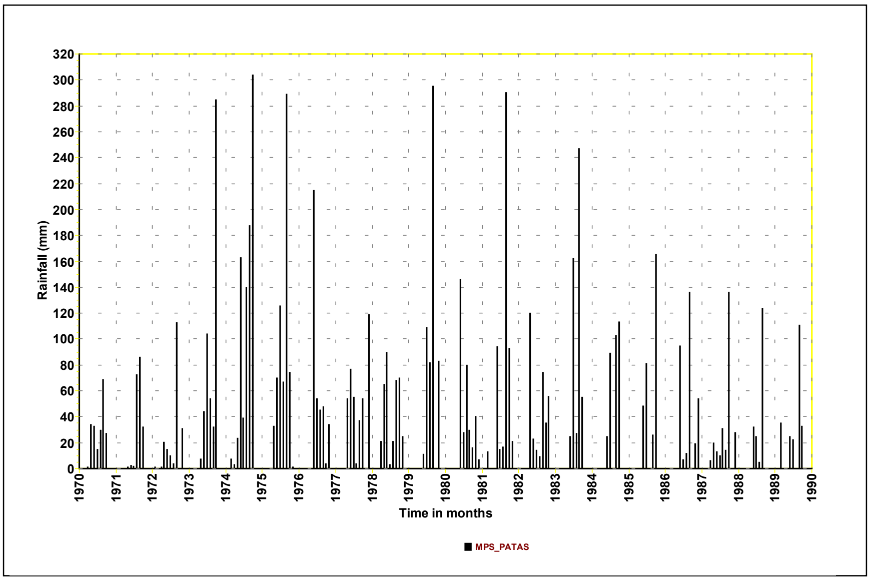

Figure 1: Monthly rainfall station PATAS, period 1979-1989

The results of the computation of the autocovariance and autocorrelation function are presented in Table 1. The autocorrelogram is shown in Figure 1.

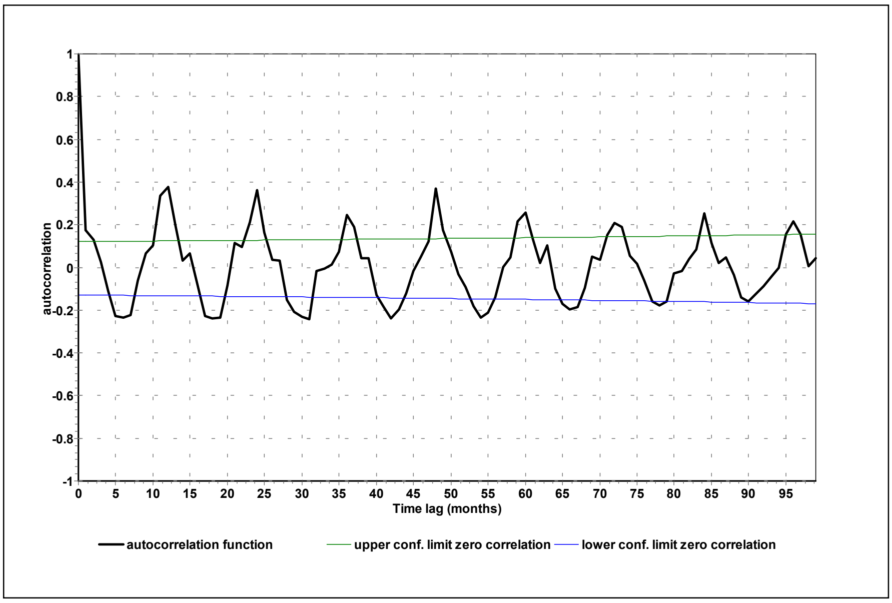

Figure 2: Autocorrelogram of monthly rainfall of station PATAS, period 1970-1989

```
Series = PATAS       MPS
Date of first element            = 1970  1  0  0  1
Date of last  element            = 1990  1  0  0  1

Confidence interval              = 95%

COV    = autocovariance function
COR    = autocorrelation function
CLP    = upper conf. limit zero correlation
CLN    = lower conf. limit zero correlation
      LAG         COV       COR       CLP       CLN
        0   .3319E+04    1.0000     .1211    -.1294
        1   .5779E+03     .1741     .1213    -.1296
        2   .4315E+03     .1300     .1216    -.1299
        3   .7955E+02     .0240     .1218    -.1302
        4  -.3789E+03    -.1142     .1221    -.1305
        5  -.7431E+03    -.2239     .1223    -.1308
        6  -.7814E+03    -.2354     .1226    -.1310
        7  -.7383E+03    -.2224     .1228    -.1313
        8  -.1996E+03    -.0601     .1231    -.1316
        9   .2174E+03     .0655     .1233    -.1319
       10   .3428E+03     .1033     .1236    -.1322
       11   .1119E+04     .3371     .1238    -.1325
       12   .1254E+04     .3779     .1241    -.1328
       13   .6556E+03     .1975     .1243    -.1331
       14   .1103E+03     .0332     .1246    -.1334
       15   .2240E+03     .0675     .1248    -.1337
       16  -.3054E+03    -.0920     .1251    -.1340
       17  -.7494E+03    -.2258     .1254    -.1343
       18  -.7914E+03    -.2385     .1256    -.1346
       19  -.7762E+03    -.2339     .1259    -.1349
       20  -.2796E+03    -.0842     .1262    -.1352
       21   .3819E+03     .1151     .1264    -.1355
       22   .3259E+03     .0982     .1267    -.1358
       23   .7051E+03     .2125     .1270    -.1361
       24   .1203E+04     .3624     .1273    -.1364 etc.
```

Table 1: Results of autocovariance and autocorrelation computation for station PATAS

From the figure and table it is observed that there is a periodicity with a period of 12 months due to the fixed occurrence of the rainfall each year in the monsoon. Any other periodicity is not apparent from the autocorrelogram. Note that the serial correlation between successive months is fairly non-existing.

The autocorrelogram for the daily series of the same station is shown in Figure 3. It is observed that there is a fast decaying correlation between rainfall on successive days. Also observe that the confidence limits for zero correlation are in this case closer to zero than for the monthly series in view of the much larger amount of data taken into the analysis (6 years of daily data).

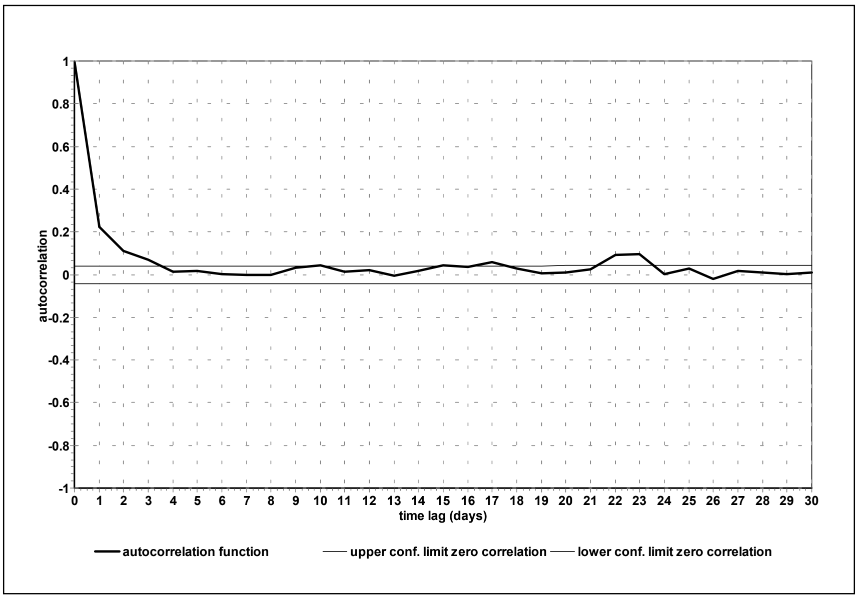

Figure 3: Autocorrelogram of daily rainfall at station PATAS (data 1990-1995)

For daily water level or discharge series the autocorrelogram will look quite different as the basin acts as a filter to the rainfall input. Then the autocorrelogram will decay much slower.

The autocorrelogram is often used to identify which time series model would be suitable to simulate the behaviour of the series. The autoregressive and moving average models do have distinct correlograms. Reference is made to textbooks on Stochastic Processes for an overview.
As an example consider a time series Y(t) of a single harmonic component, see also Figure 4:

$$\left.\begin{aligned}&Y(t) = A\sin(\omega t + \varphi) \qquad \text{or with}: \omega = 2\pi f \;\; \text{and} \quad f = \frac{1}{T}: \\ &Y(t) = A\sin(2\pi ft + \varphi) = A\sin(2\pi\frac{t}{T} + \varphi)\end{aligned}\right\} \tag{7}$$

where: A = amplitude

ω = angular frequency in radians per unit of time

f = ordinary frequency in cycles or harmonic periods per unit of time

φ =phase angle with respect to time origin or phase shift

T = period of harmonic

The covariance of a single harmonic process is derived from:

$$\left.\begin{aligned}&C_{YY}(\tau) = E[(Y(t)-\mu_Y)(Y(t+\tau)-\mu_Y)] = E[A\sin(2\pi ft+\varphi).A\sin(2\pi f(t+\tau)+\varphi)] \\ &\text{or with}: \quad \sin(a).\sin(b) = \frac{1}{2}\{\cos(a-b)-\cos(a+b)\} \quad \text{it follows}: \\ &C_{YY}(\tau) = \frac{1}{2}A^2\{E[\cos(2\pi f\tau)] - E[\cos(2\pi ft(2+\tau)+2\varphi)]\} \\ &C_{YY}(\tau) = \frac{1}{2}A^2\cos(2\pi f\tau) \qquad \text{so}: \quad C_{YY}(0) = \sigma_Y^2 = \frac{1}{2}A^2\end{aligned}\right\} \tag{8}$$

The covariance function is shown in Figure 5 together with the autocorrelation function.

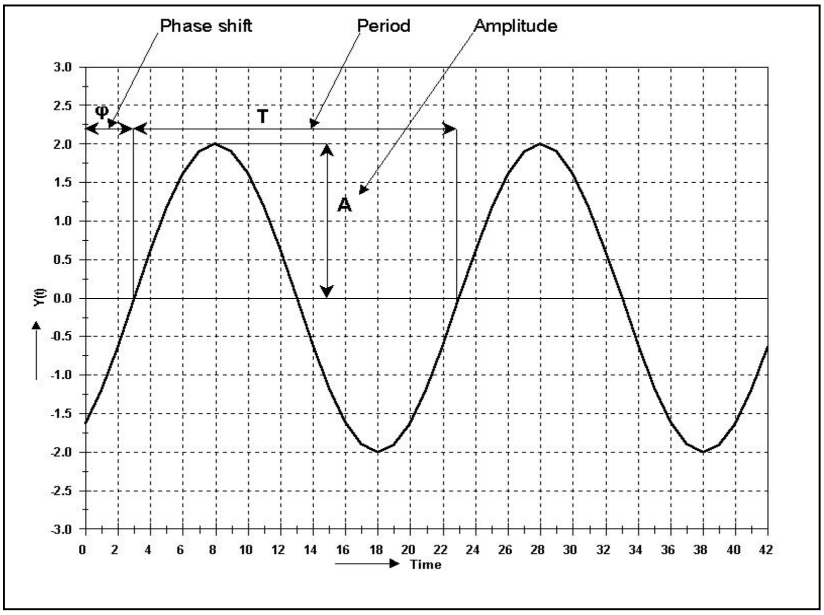

Figure 4: Series of a single harmonic

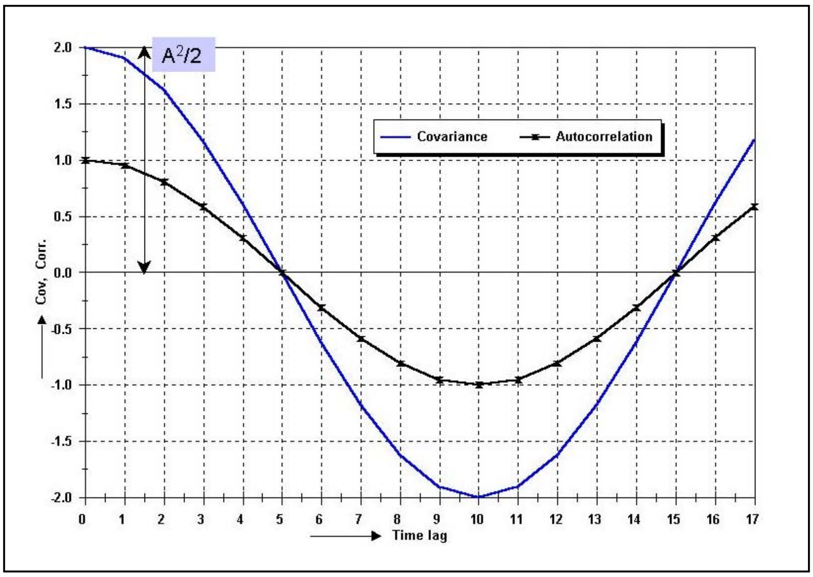

Figure 5: Covariance and autocorrelogram of single harmonic

Hence, it is observed that:

- the covariance of a harmonic remains a harmonic of the same frequency f, so the frequency information in the original series is preserved in the covariance function and the auto-correlation function.
- information about the phase shift has vanished in the covariance and correlation function.
- The amplitude of the periodic covariance function is seen to be equal to the variance of Y(t), see Figure 5.

3 Cross-Covariance and Cross-Correlation function

Similar to the way the autocovariance function of a series Y(t) was defined one can define a measure for the covariance/correlation between the elements of two series X(t) and Y(t) at times t and t+k:

$\gamma_{XY}(k) = \text{Cov}[X(t), Y(t+k)\} = E[(X(t) - \mu_X)(Y(t+k) - \mu_Y)]$ (9)

where: $\gamma_{XY}(k)$ = cross-covariance at lag k
μ_X = mean of series X
μ_Y = mean of series Y

The cross-correlation function is obtained by dividing the cross-covariance by the standard deviations of X and Y respectively σ_X and σ_Y. The graphical presentation is generally known as the cross-correlogram:

$\rho_{XY}(k) = \gamma_{XY}(k))/\sigma_X \sigma_Y$ with: $-1 \leq \rho_{XY}(k) \leq 1$ (10)

Now it is observed that for lag zero the cross-correlation is generally < 1, unless there is perfect correlation between X and Y at lag zero. The maximum may occur at some other lag, where the linear correlation between X and Y is maximum.

For example consider the hourly water level series at two stations along the same river as was shown in Module 23 (secondary validation of water level data). Dependent on the distance between the two stations and the flood wave celerity the hydrograph of the upstream station should be shifted forward in time to create the best resemblance with the hydrograph at the downstream site. Likely the cross-correlation will be maximum at a lag equal to the travel time of the flood wave between these two stations.

In HYMOS the cross-covariance function $\gamma_{XY}(k)$ is estimated by the following estimator $C_{XY}(k)$:

$$C_{XY}(k) = \frac{1}{n}\sum_{i=1}^{n-k}(x_i - m_x)(y_{i+k} - m_y) \qquad (11)$$

and the cross-correlation function $\rho_{XY}(k)$ by $r_{XY}(k)$:

$$r_{XY}(k) = \frac{C_{XY}(k)}{\sqrt{C_{XX}(0)C_{YY}(0)}} = \frac{\frac{1}{n}\sum_{i=1}^{n-k}(x_i - m_x)(y_{i+k} - m_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - m_y)^2}} \qquad (12)$$

Note that the estimators (11) and (12) are biased estimators, since the sum is divided by n rather than by (n-k-1) to obtain estimates with a smaller mean square error.

4 Variance Spectrum and Spectral Density Function

The plot of the variance of the harmonics a series can be thought of to consist of against their frequencies is called power or variance spectrum. Generally, a large number of harmonics is required to describe a process. Let $S_p(f)$ be the ordinate of the continuous spectrum, then the variance contributed by all frequencies in the frequency interval df is given by $S_p(f)$.df. Hence the total variance is obtained by integration over the full range of frequencies:

$$\int_0^\infty S_p(f).df = \sigma_Y^2 \qquad (13)$$

Since hydrological processes are frequency limited, i.e. harmonics with a frequency higher than some limiting frequency f_c do not contribute significantly to the variance of the process and can hence be eliminated. Then, approximately the following adaptation of (13) is assumed to be valid

$$\int_0^{f_c} S_p(f).df = \sigma_Y^2 \tag{14}$$

where f_c is called the Nyquist or cut-off frequency.

Generally, the spectrum is scaled by the variance (like the covariance function) to make spectra of processes with different scales comparable. It then follows:

$$S_d(f) = \frac{S_p(f)}{\sigma_Y^2} \qquad \text{hence :} \quad \int_0^{f_c} S_d(f).df = 1 \tag{15}$$

where: $S_d(f)$ = spectral density function

The spectral density function is the Fourier transform of the auto-correlation function and can be computed from:

$$S_d(f) = 2\left[1 + 2\sum_{k=1}^{\infty} \rho_{YY}(k)\cos(2\pi fk)\right] \tag{16}$$

To arrive at an estimate for the spectrum, first, $\rho_{YY}(k)$ is to be replaced by its sample estimate $r_{YY}(k)$, with k = 1,2,..,M, where M = maximum lag up to which the correlation function is estimated. It will be shown that M is to be carefully selected.

Estimating $S_d(f)$ by just replacing the auto-correlation function by its sample estimate creates a spectral estimate which has a large sampling variance. To reduce this variance a smoothing function is to be applied. With this smoothing function the spectral density at frequency f_k is estimated as a weighted average of the spectral density at surrounding frequencies e.g. f_{k-1}, f_k and f_{k+1}. This smoothing function is called a spectral window, which dimension has to be carefully designed. An appropriate window for hydrological time series is the Tukey window, which has the following form in the time domain:

$$\begin{aligned} T_w(k) &= \frac{1}{2}\left(1 + \cos(\frac{\pi k}{M})\right) && \text{for :} \quad k \le M \\ T_w(k) &= 0 && \text{for :} \quad k > M \end{aligned} \tag{17}$$

The Tukey window has a bandwidth B (indicative for the width over which smoothing takes place in the spectrum) and associated number of degrees of freedom n of:

$$B = \frac{4}{3M\Delta t} \qquad \text{and :} \quad n = \frac{8N}{3M} \tag{18}$$

The smoothed spectral estimate then reads:

$$s(f) = 2\left[1 + 2\sum_{k=1}^{M-1} T_w(k) r_{YY}(k)\cos(2\pi fk)\right] \qquad \text{for :} \quad f = 0,....,\frac{1}{2} \tag{19}$$

where: s(f) = smoothed estimator for $S_d(f)$

It is often mentioned that the spectrum is to be computed for the following frequencies (Haan (1977):

$$f_k = \frac{kf_c}{M} \qquad \text{for :} \quad k = 0,1,......,M \tag{20}$$

Jenkins and Watts (1968) argue that the spacing following from (20) is too wide and suggest that a number of frequency points equal to 2 to 3 times (M+1) is more appropriate.

A (1-α)100% confidence interval for $S_d(f)$ is obtained from:

$$\frac{n.s(f)}{\chi_n^2(1-\frac{\alpha}{2})} \le S_d(f) \le \frac{n.s(f)}{\chi_n^2(\frac{\alpha}{2})} \tag{21}$$

In Figure (6) the 95% confidence limits are shown for s(f) = 2. From Figure 6 it is observed, that the variance of the spectral estimate reduces with increasing number of degrees of freedom, i.e. according to (18) with decreasing number of lags M in the computation of the auto-correlation function.

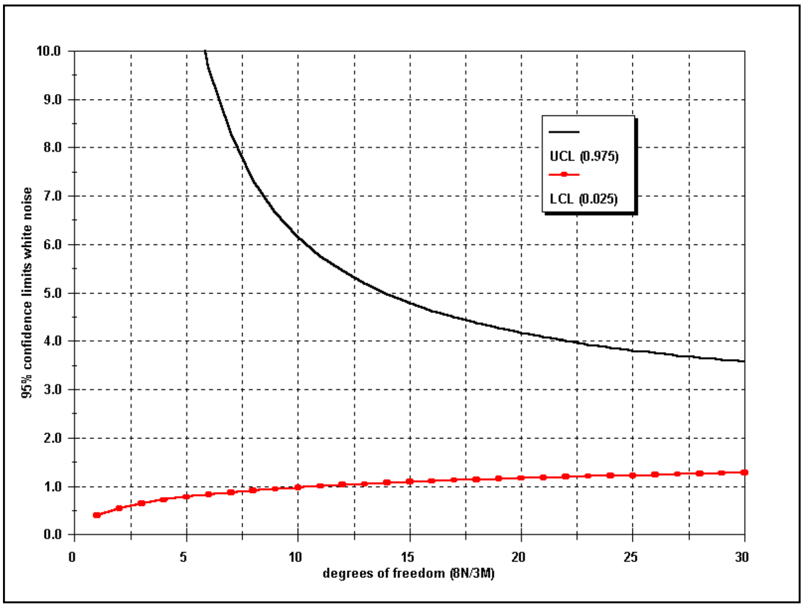

Figure 6:
95% confidence limits for white noise

Hence, to reduce the sampling variance the maximum lag M should be taken small. From Figure 6 it is observed that for say n > 25 the sampling variance reduces only slowly, so little further improvement in the estimate is obtained beyond that point. Consequently, a value for M of about 10 to 15% of N will do in line with (18). But a small value of M leads, according to (18), to a large bandwidth B. The value of B should be smaller than the frequency difference between two successive significant harmonics in the spectrum. Say a water level is sampled at hourly intervals, hence Δt = 1 hour. One expects significant harmonics with periods of 16 and 24 hours. The frequency difference between the two is 1/16 - 1/24 = 1/48. Hence, it is requested that: B < 1/48, so the condition for M becomes: M > 4x48/3 = 64. If one chooses M to be 10% of N then at least 640 data points should be available for the analysis, i.e. some 27 days or about one month of hourly data. When a series of given length N is available, it follows for the range of acceptable values of M: $4/(3B) < M \le 8N/(3n)$ where one should not be too close to the lower level, and an acceptable value $n \approx 25$. It is to be noted though that since it is not known in advance which harmonics will be significant, that this process is repeated for different values of M.

To investigate which harmonics are significant its variance should be outside the confidence limits for white noise. A white noise process means a random process without any correlation between successive data points. According to (16) with $\rho_{YY}(k) = 0$ for $k > 0$ it follows that $S_d(f) = 2$ for $0 \leq f \leq ½$ The confidence limits for a white noise spectrum are obtained from (21) by substituting for s(f) the average spectral density estimated for $0 \leq f \leq ½$. This average should be ≈ 2. The 95% confidence levels are shown in Figure 6.

Example 2

The monthly rainfall data series of station PATAS have also been subjected to spectral analysis. The results are shown in Figure 7 and Table 2

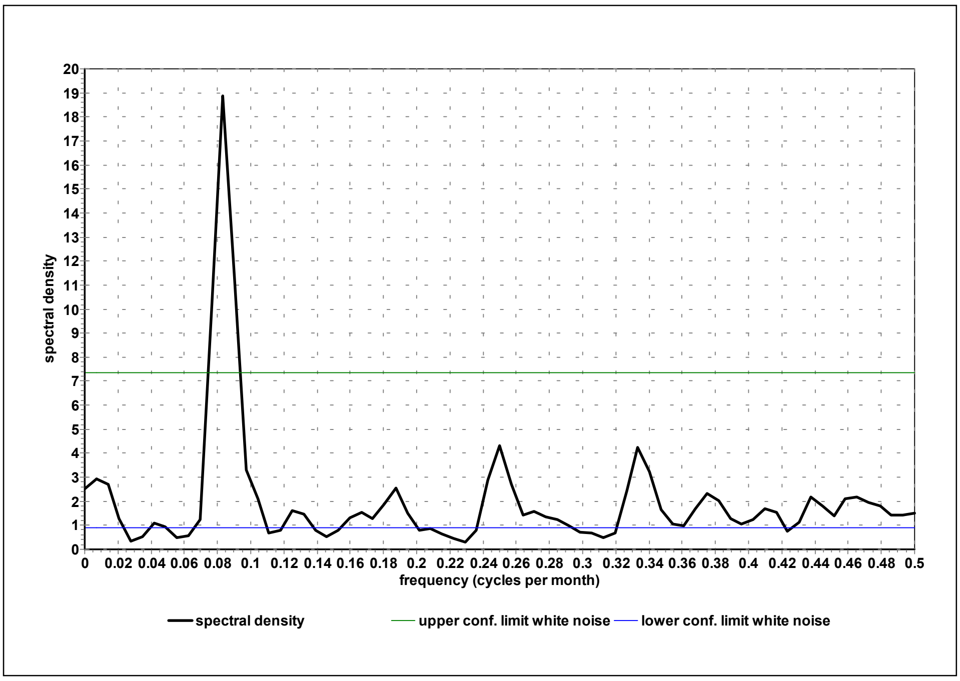

Figure 7: Spectral density function for PATAS monthly rainfall data

```
Series     =PATAS       MPS
Date of first element            = 1970  1  0  0  1
Date of last  element            = 1989 12  0  0  1

Truncation lag               =    72
Number of frequency points   =    72

Bandwith                     =     .0185
Degr.frdom                   =     8

upper conf. limit white noise =     .9125
lower conf. limit white noise =    7.3402

ASPEC      = variance spectrum
LOG SPEC   = logarithm of ASPEC
DSPEC      = spectral density
```

```
NR FREQUENCY       ASPEC   LOGSPEC      DSPEC
 0     .0000   .8379E+04    3.9232     2.5174
 1     .0069   .9706E+04    3.9870     2.9162
 2     .0139   .8942E+04    3.9514     2.6866
 3     .0208   .4188E+04    3.6220     1.2583
 4     .0278   .1125E+04    3.0511      .3380
 5     .0347   .1776E+04    3.2495      .5337
 6     .0417   .3632E+04    3.5601     1.0912
 7     .0486   .3158E+04    3.4993      .9487
 8     .0556   .1659E+04    3.2199      .4984
 9     .0625   .1861E+04    3.2698      .5592
10     .0694   .4128E+04    3.6157     1.2401
11     .0764   .3419E+05    4.5339    10.2732
12     .0833   .6287E+05    4.7985    18.8906
13     .0903   .3767E+05    4.5759    11.3166
14     .0972   .1104E+05    4.0430     3.3173
15     .1042   .7031E+04    3.8470     2.1124
16     .1111   .2293E+04    3.3604      .6890
17     .1181   .2650E+04    3.4233      .7962
18     .1250   .5415E+04    3.7336     1.6269
19     .1319   .4855E+04    3.6862     1.4587
20     .1389   .2575E+04    3.4108      .7737 etc.
```

Table 2: Results of spectral analysis of monthly rainfall data for station PATAS

From figure and table it is observed that maximum variance is available for f = 0.0833 = 1/12 cycles per month. It confirms the observation made in Chapter 2 from the autocorrelogram that a strong annual cycle is present in the series.

Nyquist or cut-off frequency

When the spectrum is made for measurements taken at small intervals Δt, and if no significant harmonics are available for f in the interval $f_c < f \leq 1/2\Delta t$, it follows that a larger sampling interval may be applied without loss of information. Note that for a harmonic component with period T the sampling interval Δt should be < 1/2T to be able to detect the harmonic (more than two samples per period). Hence, it follows that when a hydrological process has a cut-off frequency f_c, this implies that harmonics with a period $T < T_c = 1/f_c$ do not contribute significantly to the series variance. By sampling this process at intervals $\Delta t < 1/2T_c = 1/(2f_c)$ apart will not lead to loss of information. So, the sampling interval Δt should fulfil the criterion:

$$\Delta t < \frac{1}{2f_c}$$

to ensure full reproduction at all characteristics of the process. A much smaller sampling interval will lead to redundancy.