TR- 147

STOCHASTIC MODELLING OF WATER QUALITY USING DATA FOR RIVER YAMUNA

NATIONAL INSTITUTE OF HYDROL
JAL VIGYAN BHAWAN
ROORKEE-247667 (U.P.)
INDIA

1992-93

PREFACE

Water is one of the most essential constituents of the human environment. As a part of the general concern for environment, water quality has become an important water resources issue due to the increasing trend of pollution sources e.g. rapid population growth, rapid industrial development, increasing mining and petroleum operations, and too much use of fertilisers and pesticides in agriculture

Water quality is usually described by set of physical, chemical, and biological parameters which are mutually inter-related and vary according to a complex function of natural and man made interactions in both time and space. Both of these interacting mechanisms affecting water qualty are to a certain extent affected by the laws of chance. This is particularly true for the effects introduced through variation in the natural hydrologic cycle. To properly interpret water quality data, it is critical that the random nature of water quality variables be appreciately understood. Emphasis should be given on understanding how water quality parameters evolve under natural and society affected conditions in various bodies of water. Furthermore, the deterministic water quality analysis is complicated by the fact that many of the factors which influence variations in water quality are still not well defined. These factors may be further obscured by the occurrence of random events. Consequently, the application of stochastic techniques to water quality data has become necessary in order to generate the data for various water quality parameters which can be used for water quality modelling studies.

In the present study, stochastic modelling technique with particular emphasis on river water quality, is described and applied, to model the mean monthly dissolved oxygen data observed at U/S and D/S sections of Yamuna river at Delhi. Data published by CWC, New Delhi in "Water Quality Studiess- Yamuna System (1978-90)" for river Yamuna were used for the study.

The study entitled "Stochastic Modelling of Water Quality

Using Data for River Yamuna" has been carried out by sh.Aditya Tyagi, scientist 'B',Environmental Hydrology Division, NIH. The scientific helps provided by Dr.S.M.Seth, scientist 'F', sh.R.D.Singh,scientist 'E', sh.Avinash Agarwal, scientist 'C', sh.N.C.Ghosh, scientist 'C' of NIH are mentionable.

Satish Chandra
(SATISH CHANDRA)

CONTENTS

ABSTRACT

In the past two decades, many mathematical water quality models have been developed to simulate physical, chemical, and biological processes occurring in river water. Their possible applications range from identifying in streaming processes affecting river water quality to forecasting the quality for operational purposes.

It was a common practice to describe problems related to chemical and biological processes in river waters through deterministic differential equations. Since the deterministic model provides a single response for each set of model parameters and initial conditions, there is always some uncertainty, both in the evaluation of field data and in the use of mathematical models to predict the outcome of natural processes. The full representation of the process responses is usually too complicated and may be too costly to develop. Due to inherent variability and randomness in natural processes and their measurements, all these sources of uncertainty could be represented as input forcing terms in the balance equations. The initial conditions may be random, either because of the imperfect real initial conditions or because of the biased measurements. The model coefficients (rate constants) may be random due to variations in measurements.

Number of models have been proposed in recent years which treat water quality processes as stochastic. In the present study,a time series analysis approach was applied to model nine years of mean monthly dissolved oxygen data observedat U/S and D/S sections in river Yamuna at Delhi. The data was measured and compiled by Central water commission, New Delhi,in a form of status report on water quality survey for the Yamuna system.

The basic properties of the water quality data time series were determined, time and frequency-domain analysis werecarried out, and the depedent stochastic component was represented by various stochastic models. The independent residualcomponent was represented by probablity distribution functions.

1.0 INTRODUCTION

Water is an essential element in the maintenance of all forms of life, and most living organisms can survive only for short periods without water. This fact has resulted in the development of direct relationship between abundance of water, population density, and aquatic life. As well as being in abundant supply, the available water must have specific characteristics. As a part of the general concern for environment, water quality become an important water resources issue due to the increasing trend of pollution sources e.g. rapid population growth, rapid industrial development, increasing mining and petroleum operations, and too much use of fertilizers and pesticides in agriculture. Hence it becomes highly essential to protect the water resources from the various types of pollution.

The proper management of water resources even on a small scale is very difficult. There are a large number of quality criteria to be considered and in most cases the level of each criteria is the complex intrections. The situation is further exacerbated by the difficulties of any experimental approach in forecasting water quality. This has led to the growth of mathematical modelling as a means of predicting quality.

The representation of the intrection in a system by set of equations is not a new idea. The classic work on oxygen sag by Streeter and Phelps demonstrated the possibilities. But until recently, the application of mathematical modelling was limited by the difficulty of finding analytical methods of solutions that has led to increasing interest in modelling.

In the past two decades, many mathematical water quality models have been developed to simulate chemical, physical, and biological processes occuring in river waters. Their applications range from identifying in streaming processes affecting river water quality to forecasting the quality for operational purposes. Broadly speaking, three kinds of mathematical approaches have beenused for the development of mathematical water quality models. They may be classified under the following headings.

1.1 Deterministic approach:

It was a common practice to describe the problems related to

chemical and biological processes in river waters through deterministic differential equations. The deterministic approach has been used for predicting the steady state water quality conditions along a river and to predict the short term transient state of water quality parameters (Falkher 1972; Dresnack and Dobbins 1968). The models were deterministic in that they provided a single response for each set of model parameters and initial conditions. The deterministic models include QUAL II, SSAM, and the DOMOD series of the Ontario Ministry of the Environment (MOE) and Weatherbe.

The deterministic modelling approach is important because it makes it possible to understand the cause and effect relationships that govern water quality in a river. Once cause and effect relationships are known, management alternatives can be explored and the result of any improvements and changes can be projected.

There is always some error or uncertainty in a model. A mathematical model can not represent the real process perfectly, either there is some unknown process involved or some part of the representation which can not be calculated due to complexity or economics.

1.2 **Stochastic approach:**

The water quality variations in a river may be modelled by the stochastic approach in which the variation of the magnitude of one or more parameters of water quality are represented as a function of time (or space). As, there is always some uncertainty, both in the evaluation of field data and in the use of mathematical models to predict the outcome of natural processes, since the processes are still not completely understood and the full representation is usually too complicated and too costly to implement. There is also some inherent variability and randomness in natural processes and their measurements. All these source of uncertainty may be represented as input forcing terms in the balance equations. The initial conditions may also be random, either because the knowledge of the real initial conditions is imperfect or because measurements are biased by random variations. The model coefficient may also be random either because our assessment is not perfect or because of random variations in

measurements. Inputs may also be uncertain because estimates of future loadings, based on projections and future waste technologies, may be biased. As a result of these factors the application of stochastic modelling approach become necessary.

A number of models had been proposed in recent years which treat water quality process as stochastic. The most common approach is based on using Monte Carlo techniques (Esen and Bennet 1971; Shih 1975; Dewey 1984). Unfortunately, very often these techniques are limited because of the time required for the computations. The other techniques include MARKOV chains, birth-death and random walk processes and ARIMA models. The present study discuss the use of ARIMA models.

1.3 Deterministic and stochastic Combined approach:

There is also a third approach in which the water quality equations have been transformed from deterministic to stochastic differential equations (Soong 1973; Leduc et.al 1986) used the Fokker-Plank equation to get the probability density function (pdf) of remaining CBOD and oxygen consumed, and used moment equations to obtain the expectation and variance of the first order CBOD equation. Finney et.al. (1982) developed the model to compute the joint and marginal pdf of CBOD and DO. In addition, moment equations were also developed which allowed the mean and variance of CBOD and DO to be calculated independently of their joint pdf. Dewey (1984) discussed the modification of the random differential equation method to include nitrogeneous oxygen demand (NOD) and the DO responses which had been generated directly by solution of the analytic equations. All rate constants and initial values of the vairables had been described by survey data mean values and estimated standard deviations or uniform distribution. Zeelinski (1988) developed a stochastic DO-CBOD-NOD model and applied to Thames River in Ontario, Canada.

As this method is based on the set of differnetial equations such as Streeter and Phelps (1925); Dobbin (1964) etc. which assumed certain simplifying assumptions like longitudinal dispersion is neglected; unfirom velocity along river section, plug flow; mixing is instantaneous and complete; DO saturation is temperature dependent only; variations in temperature and sunlight

negiected etc. are the known source of error.

It is concluded that the water quality variation in a river may be best modelled by the stochastic approach as the deterministic and partial deterministic approaches are not suitable due to the inherent randomness exhibited and an imposing number of uncertainties which are associated with the various processes occuring with in the stream environment.

2.0 LITERATURE REVIEW:

A time series exists as a set of observations that are, statistically, sequentially dependent. The overall aim of the analysis is to specify the character of this dependence. Time series analysis is applicable to water quality data since these data frequently exhibit such dependence. The information about periodicities which the analysis provides can suggest natural or man influenced factors that may influencing the aquatic environment. A knowledge of the processes may lead to methods of controlling water quality. The two methods of time series anlaysis, the frequency based and the time-based approach, are related in that the variance spectrum function employed in the first method is a mathematical transofmration of the autocorrelation function employed in the second. Each method has situations in which it is the best suited for analysis. The frequency approach attempts to decompose a series into its frequency components. By so doing it identifies frequencies which may then be related to factors that cause the series to vary. The method also provides an estimate of the variance attributable to each of these factors. The frequency approach has been, traditionally, the one used in water resources engineering. It has been employed e.g., to examine temperature and dissolved oxygen variations in the Delawane Estuary (Thomann, 1967), the hydraulic behaviour of charleston Harbour (Wastter et.al. 1968), and variations in waste treatment plant performances (Thomann, 1970) the time based approach of time series analysis developed by Box and Jenkins (1976) attempts to fit a model by expressing the time series as a output from a linear filter having a random input and consisting of several transfer functions in series. The method uses a minimum number of parameters.

Early applications of the Box-Jenkins method to the water resources area were undertaken by Carlson and Co-workers (Carlson et.al, 1970). They developed parametric models for annual stream flow data and were able to achieve significant reductions in variance with one or two parameters. In addition they employed the models for forecasting. McMichael and Hunter (1972) developed models for temperature and flow in rivers. Their models

incorporated both deterministic and stochastic components, the later being obtained by the Box-Jenkins method. They found this type of model to be preferable from a numerical and a rational point of view to a purely stochastic or purely deterministic model. McMichael and Vigani (1972) applied Box-Jenkins techniques to municipal treatment system organic loadings in examining a paper by Wallace and Zollman (1971). They fitted models to the authors data and then employed the models for forecasting. McKerchar and Delleur (1974) developed a model and used for forecasting the monthly steamflow by a multiplicative seasonal ARIMA model. Gupta and Chauhan (1986) developed a model for weekly irrigation requirements. Huck and Farquhar (1974) analyzed the hourly chloride and dissolved oxygen data. Lohani and Wang (1987) used the time domain analysis combined with non parametric transformation to analyze monthly water quality data in the Chung Kang river in Taiwan. More recently, Jayawardena et.al., (1989) used both time time-and frequency-domain analysis to model 21 years mean monthly water quality data in the Guangzhou reach of the Pearl river in Southern China. In which the basic properties of the water quality data time series were determined and the dependent stochastic component was represented by various stochastic models. Synthetic water quality data were generated by using the probability distribution of the independent residuals, and forecasting of future water quality data was done using a Box-Jenkins type difference model.

3.0 METHOD OF ANALYSIS

Any time series can be expressed as a linear combination of a trend component, a periodic component, and an independent residue component in the form

Time series = trend component +periodic component + dependent stochastic component + independent residue component (1)

When the components are nonlinearly related, the relationship can often be made linear by taking logarithms.Time series analyses involves the decomposition of the series in to constituent components.

A series may be stationary or nonstationary. Some nonstationary series may be made stationary by suitable treatment.

Preliminary Tests:

3.1 Trend analysis:

A steady and regular movement in a time series through which the values are, on average, either increasing or decreasing is termed a trend. This type of behavior can be local, in which case the nature of the trend is subject to change over short intervals of time, or, on the other hand, we can visualize a global trend that is long lasting. Long term trends are more appropriate to the study of hydrological time series.

3.1.1 Tests for detection of trend:

A number of tests exist for the detection of a trend, e.g.,the turning point test,Kendall s rank correlation test (Kottegoda 1980), and regression test for linear trend.

i) Turning point test

In an observed sequence x_t, t=1,2,3....,N, a turning point or p occurs at time t=i if x_i is either greater than x_{i-1} and x_{i+1} or less than the two adjacent values. The number of turning points p in a series is expressed as a standard normal variate in the form:

$$z = \frac{p - \bar{p}}{\sqrt{Var(p)}} \qquad (2)$$

where $\bar{p}$ = the expected number of turning points in a random series

$$= \frac{2(N-2)}{3};$$

$Var(\bar{p})$ = the variance of p

$$= \frac{(16N-29)}{90};$$

N = the number of observations.

ii) Kendall's rank correlation test:

This test is also based on the proportionate number of subsequent observations which exceed a particular value.For a sequence $x_1, x_2, \ldots, x_N$, The standard procedure is to determine the number of times, p, in all pairs of observations $(x_i, x_j; j > i)$ that x_j is greater than x_i ; the ordered(i,j) subsets are (i=1,j =2,3,4,.....,N),(i=2,j=3,4,5,....,N),...,(I=N-1,j=N). The test is carried out using the statistic τ defined as:

$$\tau = \frac{4p}{N(N-1)} - 1 \qquad (3)$$

The statistic is then expressed as a standard normal variate in form:

$$z = \frac{\tau - \bar{\tau}}{\sqrt{Var(\tau)}} \qquad (4)$$

where $\bar{\tau}$= the expected number of τ if the series is random (0,if random);

and $var(\bar{\tau})$ = its variance

$$= \frac{2(2N+5)}{9N(N-1)}$$

The computed standard normal variate is then compared with the standard normal variates from published tables at a given level of significance. If the calculated value of z is within the region of acceptance, the hypothesis of no trend is accepted.If a trend is detected, it can be removed by fitting a regression equation.An approximate model for describing trend is the polynomial type

$$X_t = x_o + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \ldots\ldots\ldots + \alpha_n t^n + \gamma \qquad (5)$$

in which γ is a residual term.

iii)Regression test for linear trend:

This is an alternative type of test to be used if it is thought that the trend is approximately linear. Standard methods of linear regression are used for the purpose.If we refer to equation (4), the hypothesis to be tested in this case is $\alpha = 0$.The first step is to estimate α and and its variance which are denoted by $\hat{\alpha}$ and σ^2_{α} respectively; the statistic $t = \hat{\alpha}/\sigma_{\alpha}$ is then tested

3.2 **Periodicity analysis:**

3.2.1 **Detection of Periodicity:**

Detection of periodicity can be made by the auto-correlation (time-domain) and/or spectral (frequency-domain)

analysis.If the series is periodic, the auto-correlogram will also be periodic. In the spectral density function, periodicity will appear as a peak at a frequency corresponding to the periodicity. The auto-correlation function and the spectral density function assuming stationarity, are given by –

$$r_k = \frac{\frac{1}{N-1}\left\{ \sum_{t=1}^{N-k} (X_t - \bar{X})(X_{t+k} - \bar{X}) \right\}}{\frac{1}{N}\left\{ \sum (X_t - \bar{X})^2 \right\}} \qquad (6)$$

and

$$G(f) = 2\,\Delta t\left[r_o + 2\sum_{K=1}^{M-1} r_K \cos(2\pi fk) + r_M \cos(2\pi fk) \right] \qquad (7)$$

where,

r_k = the serial auto correlation coeff. at lag k;

X_t = the observation at time t ;

G(f) = the raw spectral density function;

f = frequency;

Δt = time interval between two observation; and

M = the maximum lag considered in the auto-correlogram

3.2.2 **Representation of periodicity:**

If periodicity exists, it can be represented by a Fourier Series. The trend, if any, is assumed to have been removed at this stage. The Fourier series representation takes the form –

$$m_\tau = \mu + \sum_{i=1}^{h} \left[A_i \cos(2\pi i\tau/p) + B_i \sin(2\pi i\tau/p) \right] \qquad (8)$$

where,

m_τ = the harmonically fitted means at period τ(τ =1,2,..p);

μ = the population mean;

h = the total number of harmonics considered(=p/2 or(p+1)/2 depending on whether p is even or odd);

p = the period; and

A_i, B_i = Fourier coefficients of i^{th} harmonic.

i = integer index identifying harmonic

It is to be noted that the period p is referred to the first harmonic. For other harmonics, the arguments of the trigonometric function in equ. 10 are $2\pi\tau/(p/i)$.

The best estimate of the Fourier coefficients can be obtained by minimizing the $\Sigma(m_t - x_\tau)^2$,as given below:

$$A_i = \frac{2}{p} \sum_{\tau=1}^{p} x_\tau \cos(2\pi i\tau/p) \quad , i=1,2,\ldots \quad ,h. \tag{9}$$

$$B_i = \frac{2}{p} \sum_{\tau=1}^{p} x_\tau \sin(2\pi i\tau/p) \quad , i=1,2,\ldots \quad ,h \tag{10}$$

$$x_\tau = \frac{p}{N} \sum_{i=1}^{N/p} x_{\tau + p(i-1)} \tag{11}$$

For monthly data p =12, and therefore h=6. But for themost practical purpose, it may not be necessary to expand the Fourier series up to the maximum number of harmonics. By examining the cumulative periodigram, it is possible to determine the relative significant of each harmonic and thus obtain the maximum number significant harmonic h^*(Salas et al. 1980). The cumulative periodigram P_j, defined in the following , will show a rapidly rising part upto h* and increase slowly thereafter upto its maximum value of unity at h.

$$P = \frac{\sum_{i=1}^{j} \left[A_i^2 + B_i^2 \right] / 2}{\frac{1}{p} \sum_{\tau=1}^{} (x_\tau - \mu)^2} \tag{12}$$

where,

i = 1 to j, in decreasing order of magnitude

μ = the estimate of μ is the mean of $\bar{x}_\tau$

Now the periodic component 'm_τ' should be deducted from the series X_t, which resulted in the following new series 'Z_t':

$$Z_t = X_t - m_\tau \tag{13}$$

where,

Z_t = data series at time t, after removal of trend and periodic components.

m_t = periodic component of series X_t

In general, time series of environmental derivation fall into one of the following four categories:

1. Time series that are composed of some periodicity, a certain

degree of randomness, plus a mean with a time trend. Series of this type might be observed in cases where stream water quality is monitored over a relatively long period of time in an area experiencing industrial development.

2. Time series that are largely periodic and may include several distinct frequencies. Stream water temperature and tidal behavior generally result in time series of this type.

3. Time series that are composed of some periodicity and some degree of randomness. An example of this type can be found in the time records of dissolved oxygen in a river or estuary.1

4. Time series that appear to be characterized almost entirely by random variation. Over a relatively short period of time, the average daily sewage flow to a waste treatment plant might yield this type of time series.

It must be emphasized that the categorization of a time series is dependent not only on the length of record but also on the particular statistic of the parameter of interest which is used. For example, although the average daily sewage flow may give a time series of type (4), the hourly flow may exhibit behavior that would assign it to type (2) or (3).

3.3 Modelling of Stochastic Component

The stochastic component of the series is obtained by substracting the periodic component defined by Fourier series from the trend free series. The remaining series may have only dependent stochastic component or independent stochastic component or both the dependent and independent stochastic components. Before the further analysis it isnecessary to test the series for dependency or independency.

3.3.1 General Steps in Model Building

The main object of Box-Jenkins analysis is to find a good model that describes how the observations in a single time series are related to each other.An ARIMA model is an algebraic statement showing how a time series variable (z_t) is related to its own past values ($z_{t-1}, z_{t-2}, z_{t-3}, \ldots\ldots$). Consider the algebraic expression:

$$Z_t = C + Z_{t-1} + a_t \qquad (14)$$

Equation (14) is an example of an ARIMA model. It says that z_t is

related to its own immediately past values (z_{t-1}). C is a constant term. Φ_1 is a fixed coefficient whose value determines the relationship between z_t and z_{t-1}. The a_t is a probablistic "shock" element.

The term C, $\Phi_1 z_{t-1}$, and a_t are each components of z_t. C is a deterministic (fixed) component, $\Phi_1 z_{t-1}$ is a probabilistic component, since its value depends in part on the value of z_{t-1}, and, a_t is a purely probalistic component. Together C and $\phi_1 z_{t-1}$ represent the predictable part of z_t while a_t is a residual element that cannot be predicted within the ARIMA model. However, the a_t term assumed to have certain statistical properties.

The process of model building development by Box and Jenkins involved three basic stages e.g.,identification, estimation, and diagnostic checking. The three stage procedure is summarized schematically in fig.1.

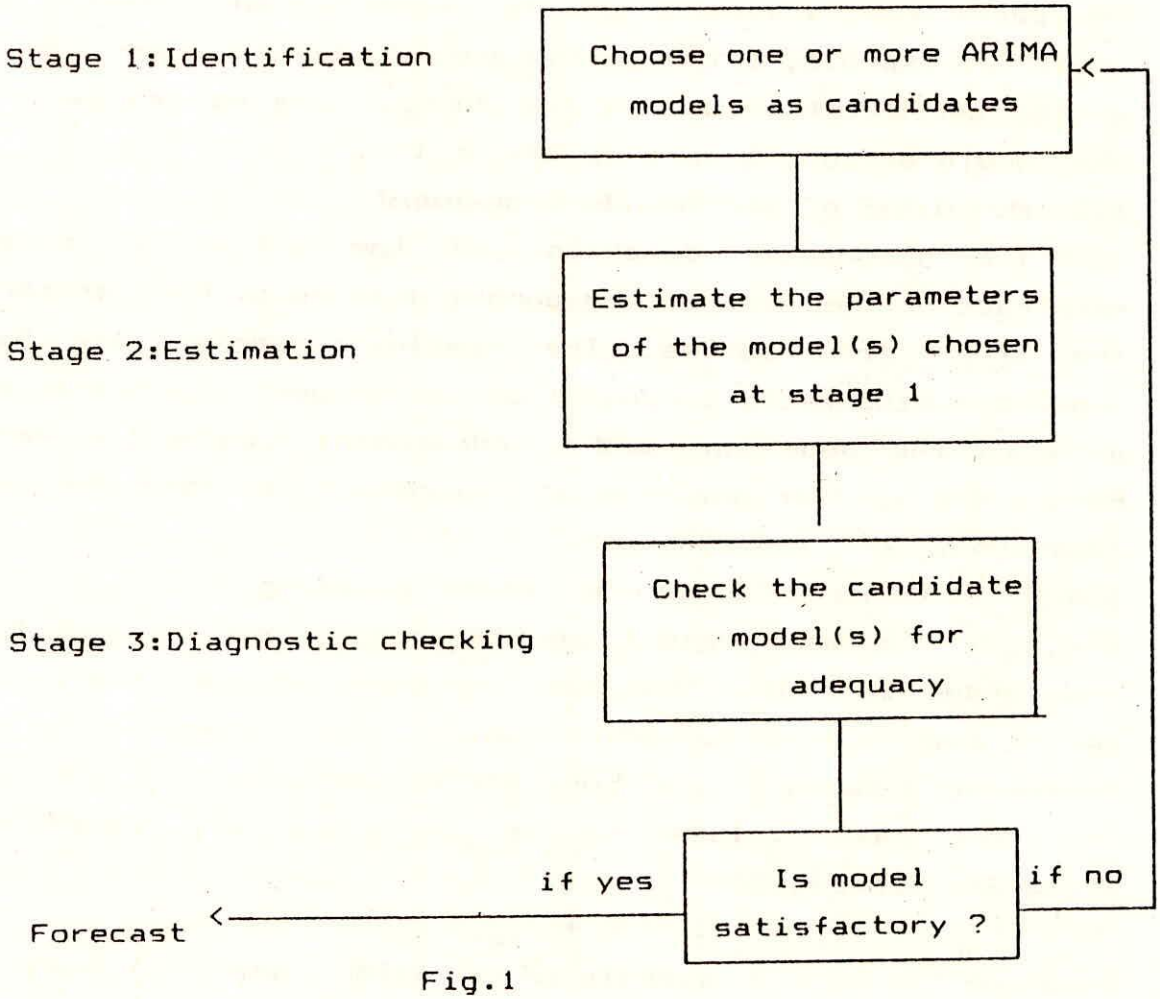

Fig.1

Stage 1: Identification

At the identification stage we use two graphical devices to measure the correlation between the observation within a single data series. These devices are called as estimated autocorrelation fuction (abbreviated acf) and an estimated partial autocorrelation function(abbreviated pacf). The estimated acf and pacf measure the statistical relationships within a data series in a somewhat crude(statistically inefficient) way. Nevertheless, they are helpful in giving us a feel for the patterns in the available data.

The next step at the identification stage is to summarize the statistical relationship within the data series in a more compact way than is done by the estimated acf and pacf. Box and Jenkins suggest a whole family of algebraic statements (ARIMA models) from which we may choose. Equation (14) is an example of such a model.

We use the estimated acf and pacf as guides in choosing one or more ARIMA models that seem appropriate. The basic idea is this: every ARIMA model, say as eqation (14), has a theoretical acf and pacf associated with it. At the identification stage we compare the estimated acf and pacf calculated from the available data with various theoretical acf's and pacf's. We then tentatively choose the model whose theoretical acf and pacf most closely resemble the estimated acf and pacf of the data series. Note that we do not approach the available data with a rigid, preconceived idea about which model we will use. Instead, we let the available data "talk to us" in the form of an estimated acf and pacf.

Which ever model we choose at the identification stage, we consider it only tentatively: it is only a candidate for the final model. To choose a final model we proceed to the next two stages and perhaps return to the identification stage if the tentatively considered model proves inadequate.

Stage 2: Estimation.

At this stage we get precise estimates of the coefficients of the model chosen at the identification stage. For example, if we tentatively choose equation (14) as our model, we fit this model

to the available data series to get estimates of ϕ_1 and C. This stage provides some warning signals about the adequacy of our model. In particular, if the estimated coefficients do not satisfy certain mathematical inequality conditions, that model is rejected.

Stage 3:Diagnostic checking

Box and Jenkins suggest some diagnostic checks to help in determining whether the estimated model is statistically adequate or not. A model that fails these diagnostic tests is rejected. Furthermore, the results at this stage may also indicate how a model could be improved. This leads us back to the identification stage. We repeat the cycle of identification, estimation, and diagnostic checking until we find a good final model. As shown in fig.1, once we find a satisfactory model we may use it for forecasting purposes.

The iterative nature of the three-stage Box-Jenkins modeling procedure is important. The estimation and diagnostic-checking stages provide warning signals telling us when, and how, a model should be reformulated, We continue to reidentify, reestimate, and recheck untill we find a model that is satisfactory according to several criteria. This iterative application of the three stages does not guarantee that we will finally arrive at the best possible ARIMA model, but it stacks the cards in our favor.

3.3.2 Analytical tools for ARIMA Modelling

The two analytical tools estimated autocorrelation function(acf) and estimated partial autocorrelation function(pacf) are very important at the identification stage of the Box-Jenkins modelling procedure. They measure the statistical relationship between observations in a single data series. These are most useful when presented in their graphical forms as wellas in their numerical forms.

i. **Estimated autocorrelation function:-**

The idea in autocorrelation analysis is to calculate a correlation coefficient for each set of ordered pairs $(\bar{z}_t, \bar{z}_{t+k})$ of the same series and the resulting statistic is called an autocorrelation coefficient which is represented by the symbol

r_k. The graphical representation of autocorrelation with the lag k is called auto correlogram.

An estimated autocorrelation coefficient (r_k) is not fundamentally different from any other sample correlation coefficient. It measures the direction and strength of the statistical relationship between ordered pairs of observations on two random variables. It is dimension less number that can take on values only between −1 and +1, value of −1 means perfect negative correlation and a value of +1 means perfect positive correlation. If $r_k=0$ then $Z_{t=k}$ and Z_t are not correlated at all in the available data.

The standard formula for calculating autocorrelation coeff. is given by equation (6). Equation (6) can also be written more compactly since $\bar{z}_t$ is defined as $(z_t-\bar{z})$, substituting accordingly and (6) becomes:

$$r_k = \frac{\sum_{t=1}^{n-k} Z_t \, Z_{t+k}}{\sum_{t=1}^{n} (Z_t)^2} \qquad (15)$$

Box and Jenkins (1976) suggest that the maximum number of useful estimated autocorrelations is roughly N/4, where N is the number of observations.

ii. **Estimated partial autocorrelation functions-**

An estimated partial autocorrelations functions (pacf) is broadly similar to an estimated acf. The estimated pacf is used as a guide along with the estimated acf in choosing one or more ARIMA models that might fit the available data.

The idea of partial autocorrelations analysis is that we want to measure how z_t and Z_{t+k} are related but with the effects of the interesting z's accounted for (i.e adjusting the impact ofany z's that fall between the ordereo pairs in question). The estimated partial autocorrelations coefficient measuring this relationship between Z_t and Z_{t+k} is desingedby statistic kk.

The most accurate way of calculating partial autocorrelation coefficient is to estimate a series of least square regression coefficient. But this mettod is complicated and require a large amount of calculation and computer memory requirement as the

number of lag increase. There is a slightly less accurate though computationally easier way to estimate the ϕ_{kk} coefficients. It involves using the previously calculated autocorrelation coeffecients(r_k).

As long as the data is stationary the following set of recrusive equations gives fairly good estimates of the partial autocorrelations.

$$\phi_{11} = r_1 \tag{16}$$

$$\phi_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\, r_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\, r_j} \tag{17}$$

$$(k = 2,3,4, \ldots\ldots\ldots)$$

where,

$$\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\, \phi_{k-1,j} \tag{18}$$

$$k = 2,3,4,\ldots\ldots, \quad j = 1,2,3,\ldots\ldots k-1.$$

For an independent series,the population correlogram is equal to zero for k=Ø. However samples of independent time series, due to sampling variability, have r_k fluctuating around zero but they are not necessarily equal to zero. In such case it is useful to determine the probability limits for the correlogram of an independent series. Anderson(1941) gave the limits-

$$r_k(95\ \%) = \frac{-1 \pm 1.96\ \sqrt{N-k-1}}{N-k} \tag{19}$$

and

$$r_k(99\ \%) = \frac{-1 \pm 1.96\ \sqrt{N-k-1}}{N-k} \tag{20}$$

for the 95 percent and 99 percent probability levels respectivaly and N is the sample size.

The another way of testing the independecy is to calculating the T-value for each r_k to measure its statistical siznificance.Any absolute t-value larger than 2 indicates that tbe corresponding r_k is significantly different from zero. The t-statistic for r_k is

$$t_{r_k} = \frac{r_k - \rho_k}{S(r_k)} \qquad (21)$$

where,

r_k = calculated value of autocorrelation at lag k

ρ_k = hypothesized value (= zero)

$S(r_k)$ = estimated standard error which is determined by the following formula

$$S(r_k) = \left[1 + 2 \sum_{j=1}^{k-1} r_j^2 \right]^{1/2} N^{-1/2} \qquad (23)$$

the t-statistic for ϕ_{kk} is-

$$t(\phi_{kk}) = \frac{\hat{\phi}_{kk} - \phi_{kk}}{S(\phi_{kk})} \qquad (24)$$

where,

$S(\phi_{kk})$ = estimated standard error which is given as-

$$S(\phi_{kk}) = N^{-1/2} \qquad (25)$$

3.3.3 Modelling of different ARIMA models and their associated charcteristics

3.3.3.1 Identification:

At the identification stage we compare the estimated acf and pacf with various theoretical acf"s and pacf"s to find a match. We choose as a tentative models from the ARIMA process whose theoritical acf and pacf best match the estimated acf and pacf. In choosing a tentative models we keep inmind theprinciple of parsimony i.e we want a models that fits the given realization with the smallest number of estimated parameters.

Table 2 state the major characteristic of theoritical acf's and pacf's for stationary AR,MA, and mixed (ARMA) process.

Table 2

Primary distinguishing characteristics of theoritical acf's and pacf's for stationary process.

Process	acf	pacf
AR	Tails of towards zero(exponential decay or damped sine wave)	cuts off to zero. (after lag p)
MA	Cuts off to zero (after lag q)	Tails off toward zero (exponential decay or damped sine wave).
ARMA	Tails off toward zero.	Tails off toward zero

The ARIMA models of higher order (e.i.,order greater than 2) do not occur often in practice. The characteristics of commonly used processes with their mathematical expressions, and their associated condition are discussed below.

AR processes:

All AR processes have theoretical acf's which tail off toward zero. This tailing off might follow a simple exponential decay pattern, a damped sine wave, or more complicated decay or wave patterns. But in all cases, there is a damping out toward zero. An AR theoritical pacf has spikes up to lag p followed by a cutoff to zero, where p is the maximum lag length for the AR terms in a process; it is also called the AR order of a process. Mathematically,the commonlly used AR processes are represented as follows:

AR(1):The common algebraic form of a stationary AR(1) process is:

$$z_t = c + \phi_1 z_{t-1} + a_t \qquad (26)$$

in backshift form this can be written as follows:

$$(1 - \phi_1 B)\ \bar{z}_t = a_t \qquad (27)$$

The estimated AR coefficients must satisfy the stationary requirement, according to which absolute value of ϕ should be less than one i.e:

$$|\phi_1| < 1 \qquad (28)$$

AR(2): The algebraic and backshift form of AR(2) process are given as:

$$z_t = c + \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t \quad (29)$$

$$(1 - \phi_1 B - \phi_2 B^2)\ \bar{z}_t = a_t \quad (30)$$

For an AR(2) process, the stationary requirement is a set of three conditions:

$$\begin{aligned} |\phi_2| &< 1 \\ \phi_2 + \phi_1 &< 1 \\ \phi_2 - \phi_1 &< 1 \end{aligned} \quad (31)$$

MA processes:

An MA process has a theoretical acf with spikes up to lag q followed by a cutoff to zero, where q is the maximum lag, also called the MA order of the process. Furthermore, an MA process has a theoretical pacf which tails off to zero after lag q. This tailing off may be either some kind of exponential decay or some type of damped wave pattern. In practice, q is usually not larger than two for nonseasonal data.The mathematical expressions for MA(1) and MA(2) processes with their invertibility conditions are given below.

The algebraic form of MA(1) and MA(2) processes are:

$$z_t = c - \theta_1 a_{t-1} + a_t \quad (32)$$

$$z_t = c - \theta_1 a_{t-1} - \theta_2 a_{t-2} + a_t \quad (33)$$

In backshift form the MA(1) and MA(2) processes can be written as:

$$(1 - \theta_1 B)\ a_t = \bar{z}_t \quad (34)$$

$$(1 - \theta_1 B - \theta_2 B^2)\ a_t = \bar{z}_t \quad (35)$$

The MA processes must satisfy the invertibility conditions which are identical to the stationary requirements on AR coefficients.

For MA(1) process, invertibility requires that the absolute value of θ_1 be less than one:

$$|\theta_1| < 1 \quad (36)$$

For MA(2) process the invertibility requirement is a set of conditions on θ_1 and θ_2:

$$\begin{aligned} |\theta| &< 1 \\ \theta_2 + \theta_1 &< 1 \\ \theta_2 - \theta_1 &< 1 \end{aligned} \quad (37)$$

ARMA processes:

Mixed processes have theoretical acf's with both AR and MA characteristics.The acf tails off toward zero after the first q-p lags with either exponential decay or a damped sine wave. The theoretical pacf tails off to zero after the first p-q lags. In practice, p and q are usually not larger than two in a mixed model for nonseasonal data.

The mathematical expressions for ARMA(1,1) and ARMA(2,2) processes are as follows:

$$z = c + \phi_1 z_{t-1} - \theta_1 a_{t-1} + a_t \qquad (38)$$

$$z = c + \phi_1 z_{t-1} - \theta_1 a_{t-1} + \phi_2 z_{t-2} - \theta_2 a_{t-2} + a_t \qquad (39)$$

The backshift form of ARMA(1,1) and ARMA(2,2) are:

$$(1 - \phi_1 B)\ \bar{z}_t = (1 - \theta_1 B)\ a_t \qquad (40)$$

$$(1 - \phi_1 B - \phi_2 B^2)\ \bar{z}_t = (1 - \theta_1 B - \theta_2 B^2)\ a_t \qquad (41)$$

The ARMA(1,q) and ARMA(2,q) processes should satisfy the stationary requirement of AR(1) and AR(2) processes respectively. Similarly, the ARMA(p,1) and ARMA(p,2) should meet the invertibility requirements of MA(1) and MA(2) processes respectively, as explained earlier.

3.3.3.2 **ESTIMATION**

At the identification stage we tentatively select one or more models that seem likely to provide parsimonious and statitically adequate representations of the available data. In making this tentative selection, a rather large number of statistics (autocorrelation and partial autocorrelation coefficients) were calculated to make proper judgement. For example, with N observations about N/4 autocorrelation and partial autocorrelation coefficient were calculated. Estimating so many parameters is not really consistent with the principle of parsimony. This nonparsimonious procedure is justifiable only as an initial, rough step in analyzing a data series. The broad overview of data contained in the estimated acf and pacf is that it gives a right direction to identify one or more appropriate

models.

ARIMA coefficients estimation can be made by three different criterion discussed as below.

i) Method of moments

ii) Method of least square

iii) Method of maximum likelihood.

When selecting an estimator or a method of estimation the two important properties should be considered. It is preferable to have both desirable properties: an unbiased estimator and a minimum mean square error(MSE) estimator. In some cases an estimator may be unbiased but it may not be minimum MSE estimator. In other cases it may be the opposite. Furthermore, estimators often are biased and do not have a menimum MSE. Therefore, when selecting among alternative estimators, a criteria is to select the estimator with the smallest bias and the smallest MSE. When this is not possible, the analyser must judge which of the two properties is more desirable for a particular case and select the estimator accordingly.

Box and Jenkins(1976) favour estimates chosen according to the maximum likelihood (ML) criterion. Mathematical statisticians frequently prefer the ML approach to estimation problems because the resulting estimates often have attractive statistical properties. However, finding exact ML estimates of ARIMA models could be cumbersome and may require relatively large amounts of computer time. For this reason, Box and Jenkins suggest the use of least squares (LS) criterion.If the random shocks are normally distributed then LS estimates are either exactly or very nearly ML estimates.

The estimation of parameters by the method of moments is usually not difficult to obtain and it is simpler than the estimation by the other methods. Except for the estimatate of the mean, the moment estimates of other parameters are usually biased, although adjustments can be applied to make them unbiased. Moment estimates are asymptotically efficient when the underlying distribution is normal. For skewed variable though, the moment estimators generally are not asymptotically efficient.

As, in the actual field problems lower order ARIMA models are

used, quite successfully. Hence in present study the ML criterion (Box-Jenkins,1970) is used for estimating the model parameters.

3.3.3.3 **DIAGNOSTIC CHECKING**

At this stage we decide if the estimated model is statistically adequate. Diagnostic checking is related to identification in two important ways. First, when diagnostic checking shows a model to be inadequate, we must return to the identification stage to tentatively select one or more other models. Second, diagnostic checking also provides clues about how an inadequate model might be reformulated.

The most important test for the statistical adequacy of an ARIMA model involves the assumption that the random shocks are independent. A statistically adequate model is one whose random shocks are statistically indepedent, meaning not autocorrelated. In practice we can not observe the random shocks (a_t), we do have the residuals ($\hat{a}_t$) calculated from the estimated model. At the diagnostic checking stage we use the residuals to test hypothesis about the independence of the random shocks.

The basic analytical tool at the diagnostic checking stage is the residual acf. A residual acf is basically the same as any other estimated acf. The only difference is that we use the residuals ($\hat{a}_t$) from an estimated model instead of the observations in a realization (z_t) to calculate the autocorrelation coefficients. To find the residual acf we use the same formula( ), but we apply it to the estimation stage residuals:

$$r(\hat{a}) = \frac{\sum_{t=1}^{n-k} (\hat{a}_t - \bar{a})(\hat{a}_{t+k} - \bar{a})}{\sum_{t=1}^{n} (\hat{a}_t - \bar{a})} \qquad (42)$$

The $\hat{a}_t$ in parentheses on the LHS of (42) indicates that we are calcualting residual autocorrelations. The idea behind the use of the residual acf is this: if the estimated model is properly formulated, then the random shocks ($\hat{a}_t$) should be uncorrelated. If the random shocks are uncorrelated, then our estimates of them ($\hat{a}_t$)should also be uncorrelated on average. Therefore, the

residual acf for a properly built ARIMA model will ideally have autocorrelation coefficients that are all statistically zero.

t-test:

Having calculated and plotted the residual autocorrelations, it is important to determine if each is significantly different from zero. The Bartlett's approximate formula,as introduced earlier in equ.(23), to estimate the standard errors of the residual autocorrelations. When applied to residual autocorrelations, the formula is:

$$s[r_k(\hat{a})] = \left[1 + 2 \sum_{j=1}^{k-1} r_j(\hat{a})^2 \right]^{1/2} N^{-1/2} \qquad (43)$$

Having found the estimated standard errors of $r_k(\hat{a})$ from equ. (43), the null hypothesis H_o: $\rho_k(a) = 0$ for each residual autocorrelation coefficient can be tested. The symbol ρ and the a in parentheses indicate that we are testing a hypothesis about the random shocks in a process. We do not have $\rho_k(a)$ values available, but we have estimates of them in the form of the residual autocorrelaTIONS $r_k(\hat{a})$. We test the null hypothesis by calculating how many standard errors (t) away from zero each residual autocorrelation coefficient falls:

$$t = \frac{r_k(\hat{a}) - 0}{s[r_k(\hat{a})]} \qquad (44)$$

In practice,if the absolute value of a residual acf t-value is less than (roughly) 1.25 at lag 1,2,and 3,and less than about 1.6 at larger lags, we conclude that the random shocks at that lag are independent. We could be wrong in this conclusion, ofcourse, but we always run that risk when making decisions based on sample information (Pankratz,1983).

If any residual acf t-value is larger than the critical value suggested above, we tentatively reject the null hypothesis and conclude that the random shocks from the estimated model are correlated and that the estimated model may be inadequate. We then tentatively identify a new model and estimate it to see if our suspicion is justified.

Chi-squared test:

This is the another way of diagonostic checking, in this the following joint null hypothesis about the correlations among the random shocks-

$$H_o: \rho_1(a) = \rho_2(a) = \ldots\ldots\ldots\ldots = \rho_K(a) = 0 \qquad (45)$$

with the test statistic

$$Q^* = N\,(N+2) \sum_{k=1}^{K} (N-k)^{-1} r_k^2(\hat{a}) \qquad (46)$$

where N is the number of observations used to estimate the model. The statistic Q^* approximately follows a chi-squared distribution with (K-m) degree of freedom, where m is the number of parameters estimated in the ARIMA model. This approximate chi-squared test is sometimes referred to as a Ljung-Box test. If Q^* is large (significantly different from zero) it says that the residual autocorrelation as a set are significantly different from zero, and the random shocks of the estimated model are probably autocorrelated. We should then consider reformulating of the model.

3.4 Modelling of Independent Stochastic Component

After indentification of the seasonol stochastic component, it was separated from the series. The new series after separation, is called as the independent stochastic component. The modelling of independent stochastic component is done by fitting the probality distributions. Now the question arises that which probability distribution should be fitted to the given data. The choice is wide, and it is likely that several probabilty distributions will fit the data equally well; the decision which to use must then be subjective. This is particularly true if the sample of data volume is small, since tests for the goodness of fit of possible distributions will have little power(i.e. these will be high probability of accepting the hypothesis that the data are consistent with the given distribution, even when this hypothesis is false).

3.4.1 Fitting a probability Distribution

A probability distribution is a function representing the probability of occurence of a random variable. By fitting a distribution to a set of water quality data, a great deal of the probabilistic information in the sample can be compactly

summarized in the function and its associated parameters. Fitting distributions can be accomplished by the method of moments or the method of maximum likelihood.

i. Method of Moments

In this method it is considered that the good estimates of the parameters of a probability distribution are those for which moments of the probability density function about the origin are equal to the corresponding moments of the sampple data. If the data values are each assigned a hypothetical mass equal to their relative frequency of occurence(1/n) and it is imagined that this system of masses is related about the origin x=0, then the first moment of each observation x_i about the origin is the product of its moment arm x_i and its mass y_n, and the sum of these moments over all the data is the sample mean, given as follows-

$$\sum \frac{x_i}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{x} \qquad (47)$$

This is equivelent to the centroid of a body. The corresponding centroid of the probability density functions is-

$$\mu = \int_{\alpha}^{\alpha} x\, f(x)\, dx \qquad (48)$$

ii. Method of Maximum Likelihood

In this method it is considered that the best value of a parameter of a probability distribution should be that value which maximizes the likelihood or joint probability of occurence of the observed sample. Suppose that the sample space is devided in to intervals of length dx and that a sample of independent and identically distributed observation $x_1, x_2, \ldots, x_n$ is taken. The value of the probability density for $X=x_i$ is $f(x_i)$, and the probability that the random variable will occur in the interval including x_i is $f(x_i)dx$. Since the observations are independent, their joint probability of occurence is given by the product $(f(x_1)\,dx)\,(f(x_2)\,dx.)\ldots(f(x_n)dx) = \left[\prod_{i=1}^{N} f(x_i)\right] dx^n$, and since the interv size dx is fixed, maximising the joint probability of the observed sample is equivalent to maximising the likelihood function-

$$L = \prod_{i=1}^{N} f(x_i) \qquad (48)$$

Because many probability density functions are exponential, it is sometimes more convenient to work with the log-likelihood function-

$$\ln(L) = \sum_{i=1}^{N} \ln [f(x_i)] \qquad (49)$$

The method of maximum likelihood is the most theoritically correct methoc of fitting probability distributions to data in the sense that it produces the most efficient parameter estimates-those which estimate the population parameters with the least average error. But, for some probability distributions, there is no analytical solution for all the parametrs in terms of sample statistics, and the log-likelihood function must then be numericallly maximized, which may be quite difficult. In genral, the method of moments is easier to apply then the method of maximum likelihood and is more suitable for practical analysis.

3.4.2 Various Probability Distributions for Hydrologic Variables

In this section, a selection of probability distributions commonly used for hydrologic variables is presented.

i. **Log-normal Distribution:-**

This has the following functional form -

$$f(y) = \frac{1}{y\,\sigma\sqrt{2\pi}} \exp\left[\frac{-(\log_e y - \mu)^2}{2\,\sigma^2}\right] dy, \qquad 0 < y < \alpha \qquad (50)$$

The properties of this distribution are-

i) that the variable $Y=\log_e y$ has a normal or ganssian distribution with mean μ and variance σ^2.

ii) that is unimodal, skewed, with a 'tail' extending to right.

ii. **The two-parameter Pearson Type III (gamma) distribution:**

This has the following functional form -

$$f(y)dy = \frac{\left(\frac{y}{\alpha}\right)^{p-1} \exp\left(-\frac{y}{\alpha}\right) dy}{\alpha\,\lceil p}, \qquad 0<y<\alpha\ ,p>1 \qquad (51)$$

Some properties of this distribution are-

i) that its shape is determined by the two parameters α and p.

ii) that its mean is αp and its variance $\alpha^2 p$.

iii) that it is unimodal for $p>1$, skewed, with a 'tail' extending to the right.

The parameters α and p may be estimated by the method of moments or by the method of maximum likelihood.

iii. **The three Parameter Pearson Type III distribution:**

This distribution has the following functional form

$$f(y)dy = \frac{1}{\sigma \lceil p} \left(\frac{y-\alpha}{\sigma}\right)^{p-1} \exp\left[-\left(\frac{y-\alpha}{\sigma}\right)\right]dy, \quad \alpha<y<\alpha, \quad \alpha>0 \quad \&p>2 \tag{52}$$

which is determined by the three parameters, p, α ,and σ. These may be estimated either by the method of moments or the method of maximum likelihood

Data source : Annex 4.7/2, status report on water quality for Yamuna system, Central water commission, New Delhi, April, 1990.

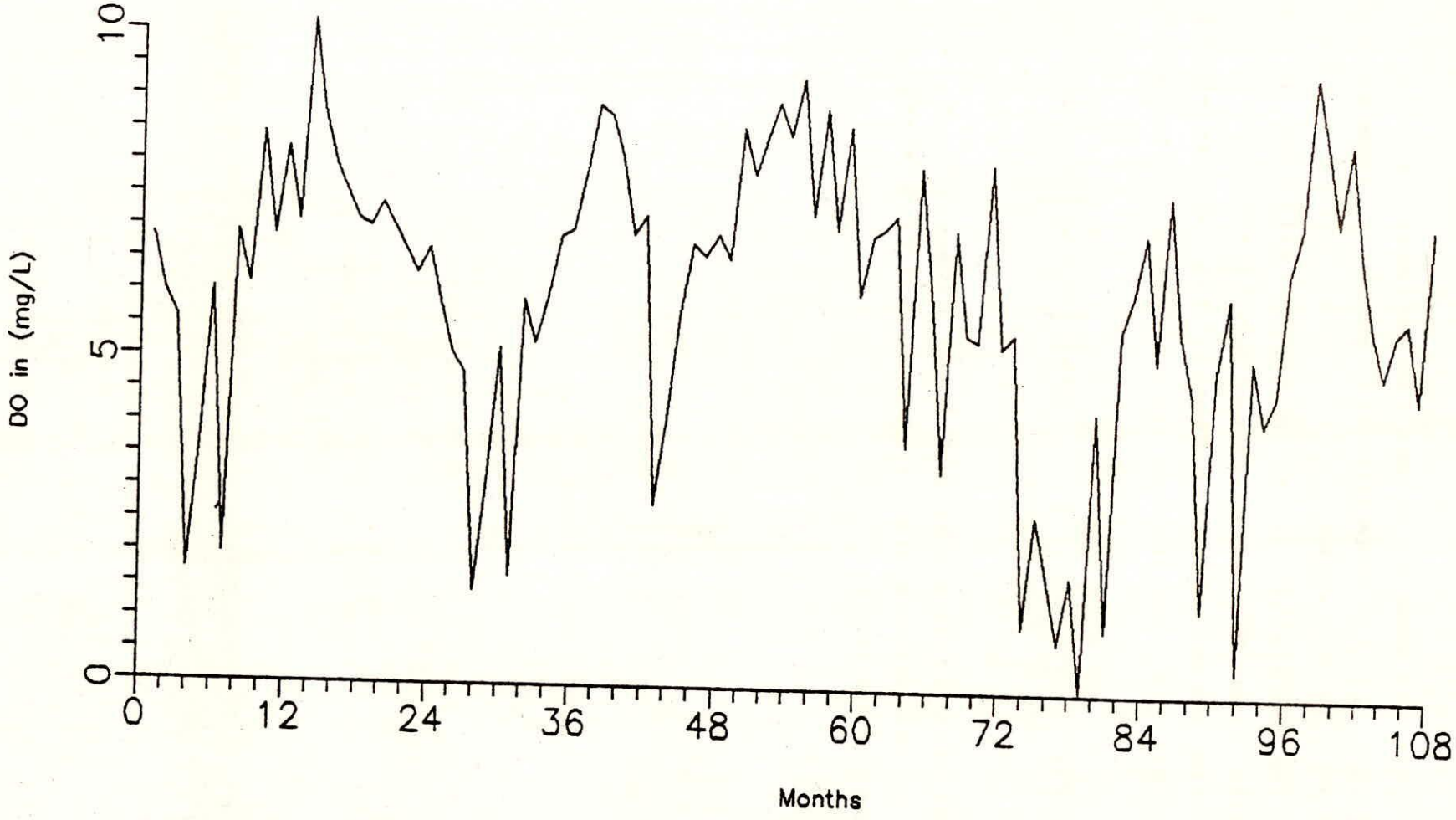

Fig. 2 ; Time series of monthly mean DO concentrations at U/S section in Yamuna at Delhi for the period Nov.,1981 – Oct.,1990.

Data source : Annex 4.7/2, status report on water quality for Yamuna system, Central water commission, New Delhi, April, 1990.

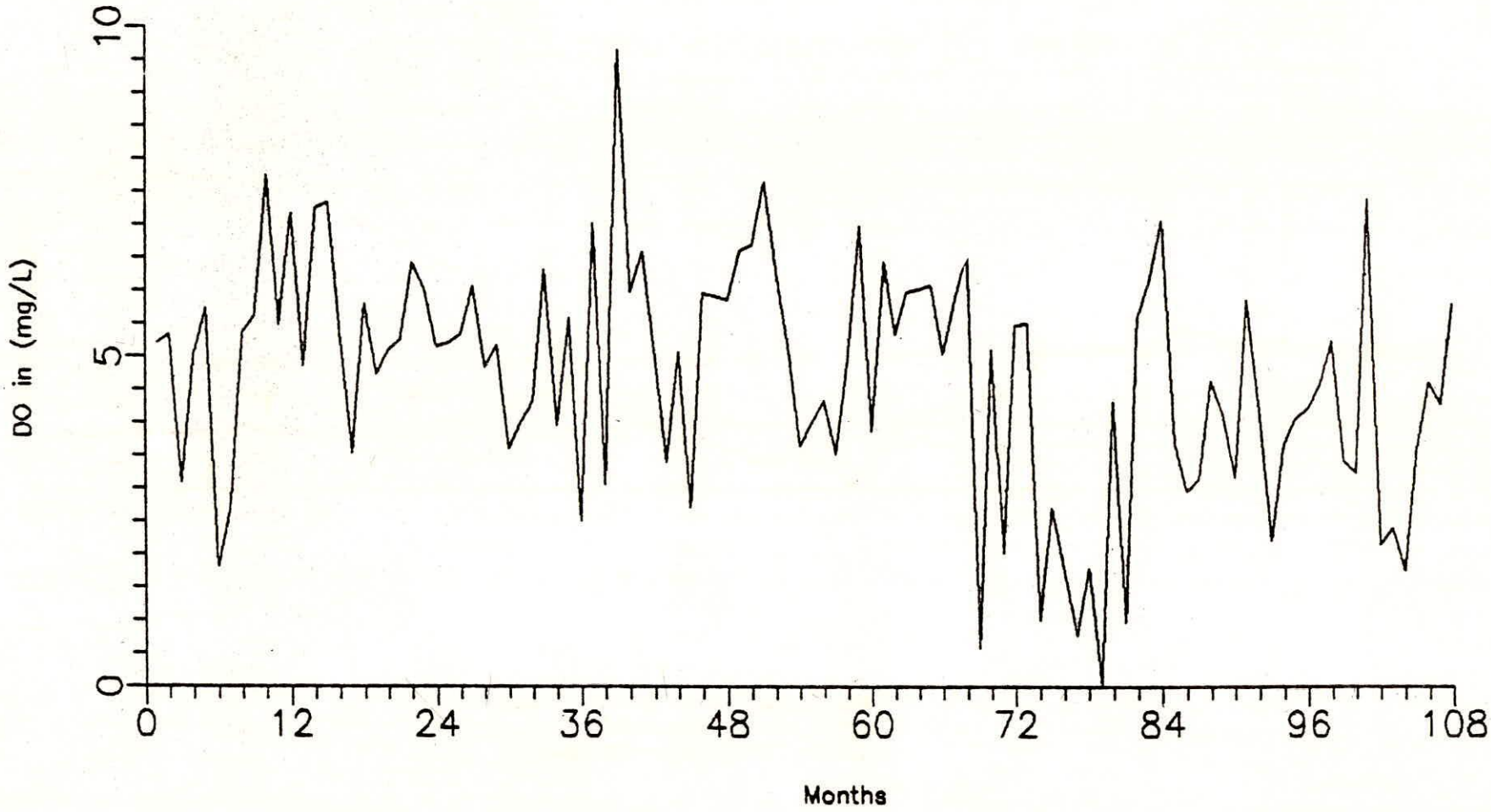

Fig. 3 : Time series of monthly mean DO concentrations at D/S section in Yamuna at Delhi for the period Nov.,1981 – Oct.,1990.

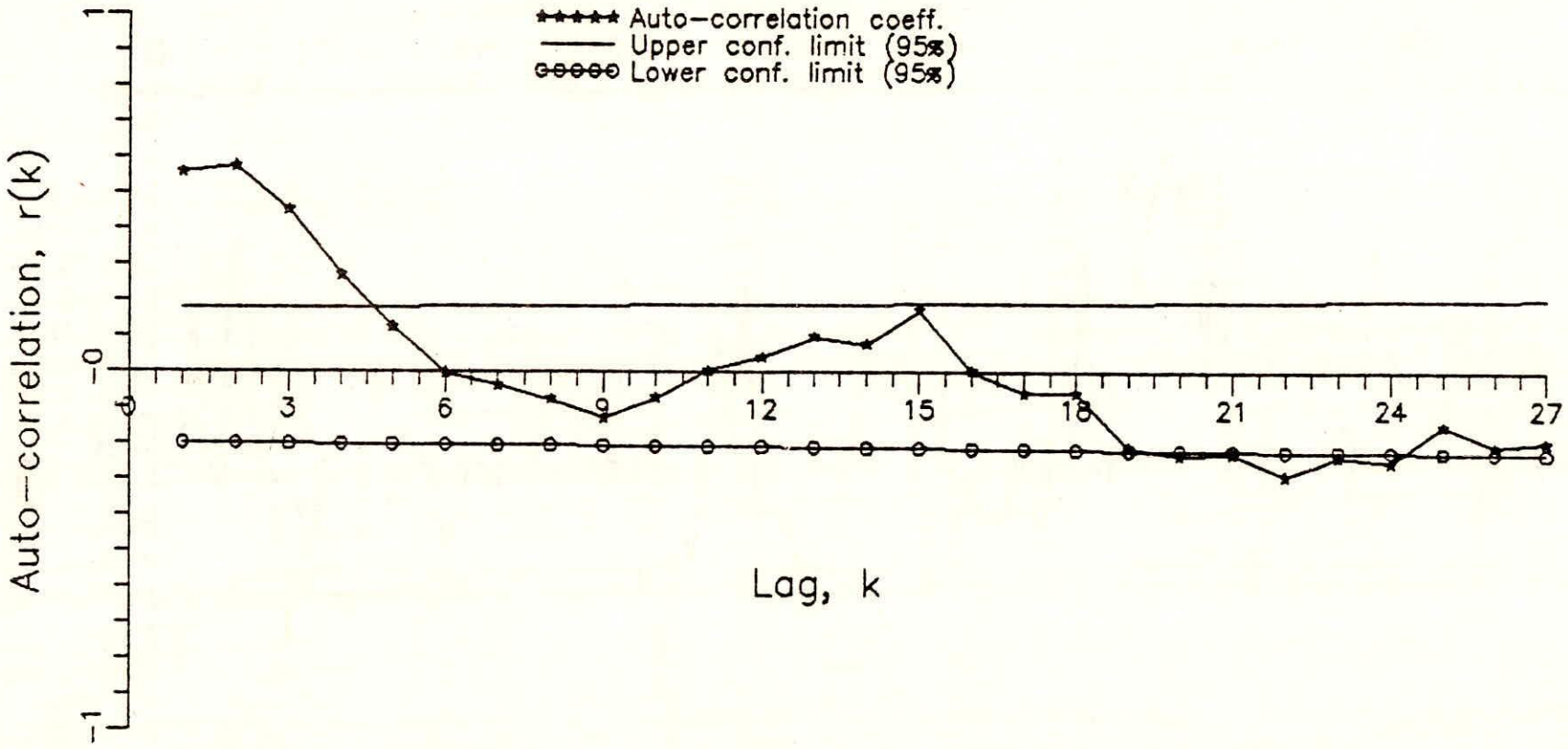

Fig. 4 : Auto-correlation function of historical DO series at U/S of Yamuna in Delhi.

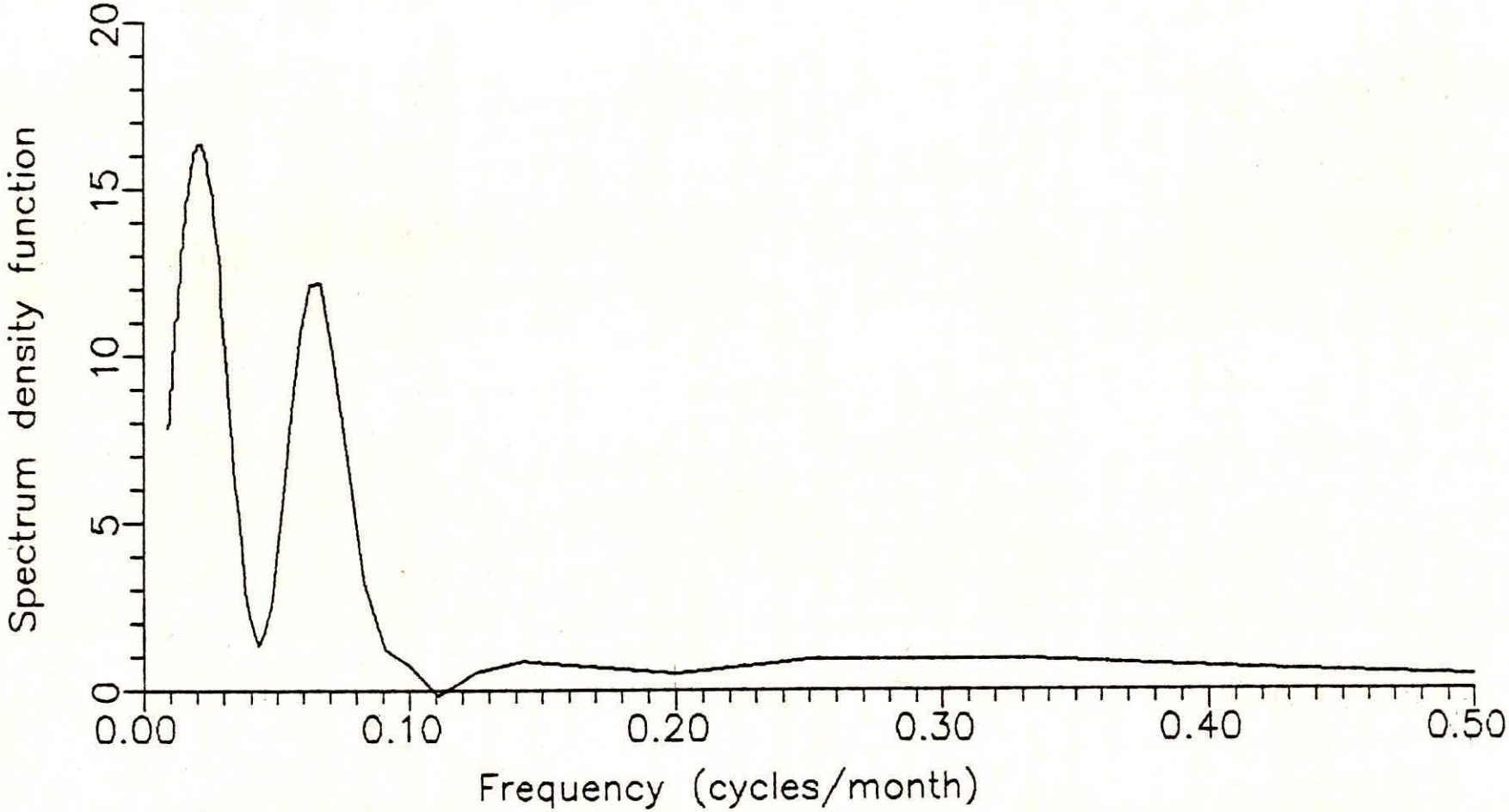

Fig. 5 : Power spectrum of U/S DO series at Delhi on river Yamuna

4.0 Data Analysis

The monthly observation for dissolved oxygen of the river Yamuna at upstream and downstream of delhi were used for the water quality analysis. Fig. 2 and fig. 3 shows the time series of mean monthly dissolved oxygen concentrations for the period 1977-1985. The data were observed and edited by the CWC.

4.1 Trend Component

The turning point test and Kandall's rank correlation test were carried out for the detection of trend. For U/S section the z value (-1.64)indicated no evidence of trend at the 5% level of significance(standard normal variate from published table at 95 % level of significance is ± 1.96). This was confirmed by tests for the detection of linear trends (t value -1.64 which is less than t-critical ± 1.98).For D/S section the z value (-3.08) indicate the existance of trend component. Further it also shows the existance of linear trend (t value -3.45).

Kottegoda (1980) suggest that only annual data should be used for the analysis of trend by virtue of which the periodic component P_t is suppressed.The Kendall's rank co-rrelation tests were carried out using annual data which indicated that there were no existance of trend component at both U/S and D/S sections of the river as the calculated z-values are less than their critical values.

For U/S section z = -0.417, and

For D/S section z = -1.67

4.2 Periodic Component:

Periodicities in the series were identified through the construction of auto-correlogram and the spectrum analysis.

For U/S section auto-correlogram (Fig.4) and the spectral density function (Fig.5) show the periodic character of the Dissolved oxygen concentration. From the spectral density function it is clear that only first two harmonics are significant for U/S section. Table 2 summarizes the results of the harmonic analysis for U/S section.

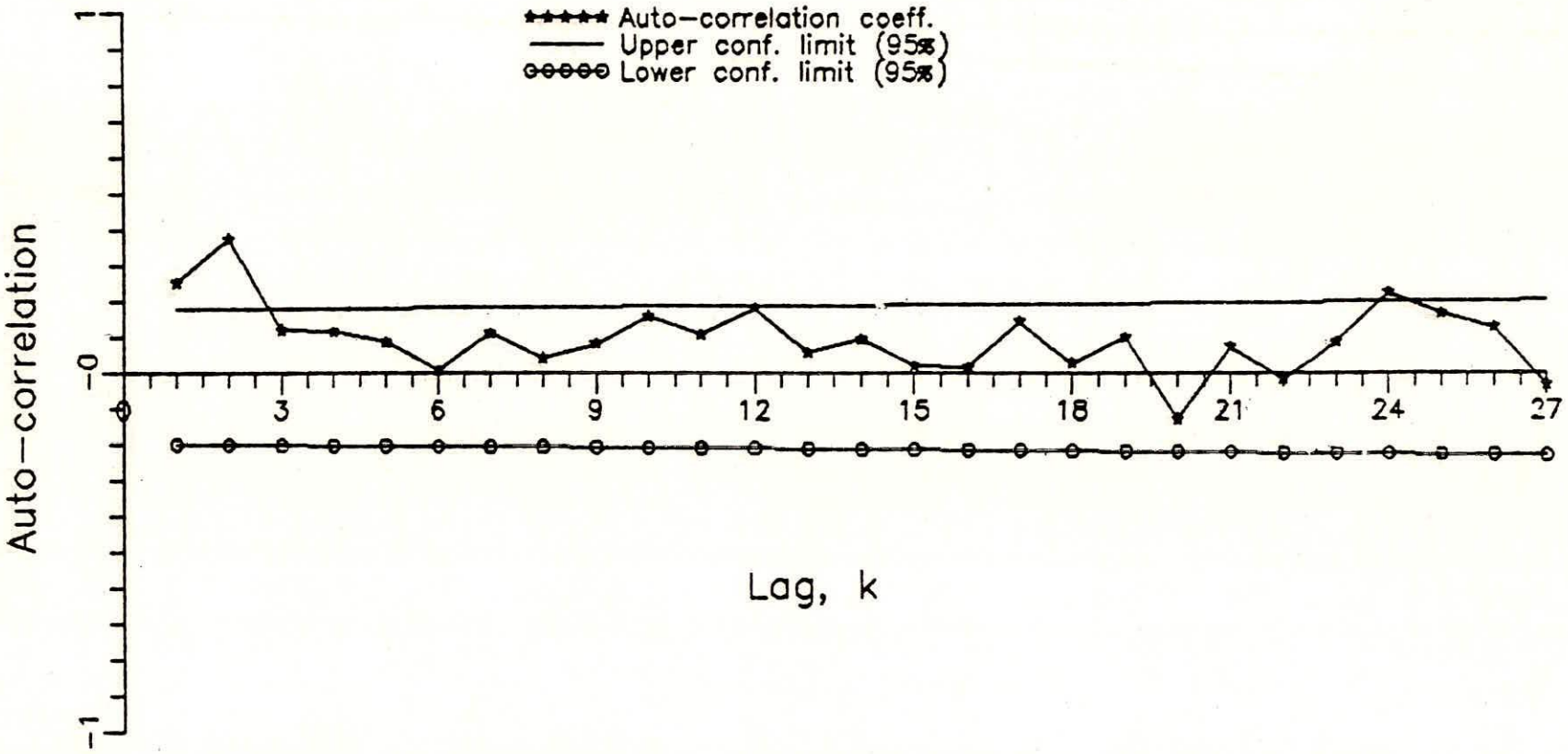

Fig. 6 : Auto–correlation function of historical DO series at D/S of Yamuna in Delhi.

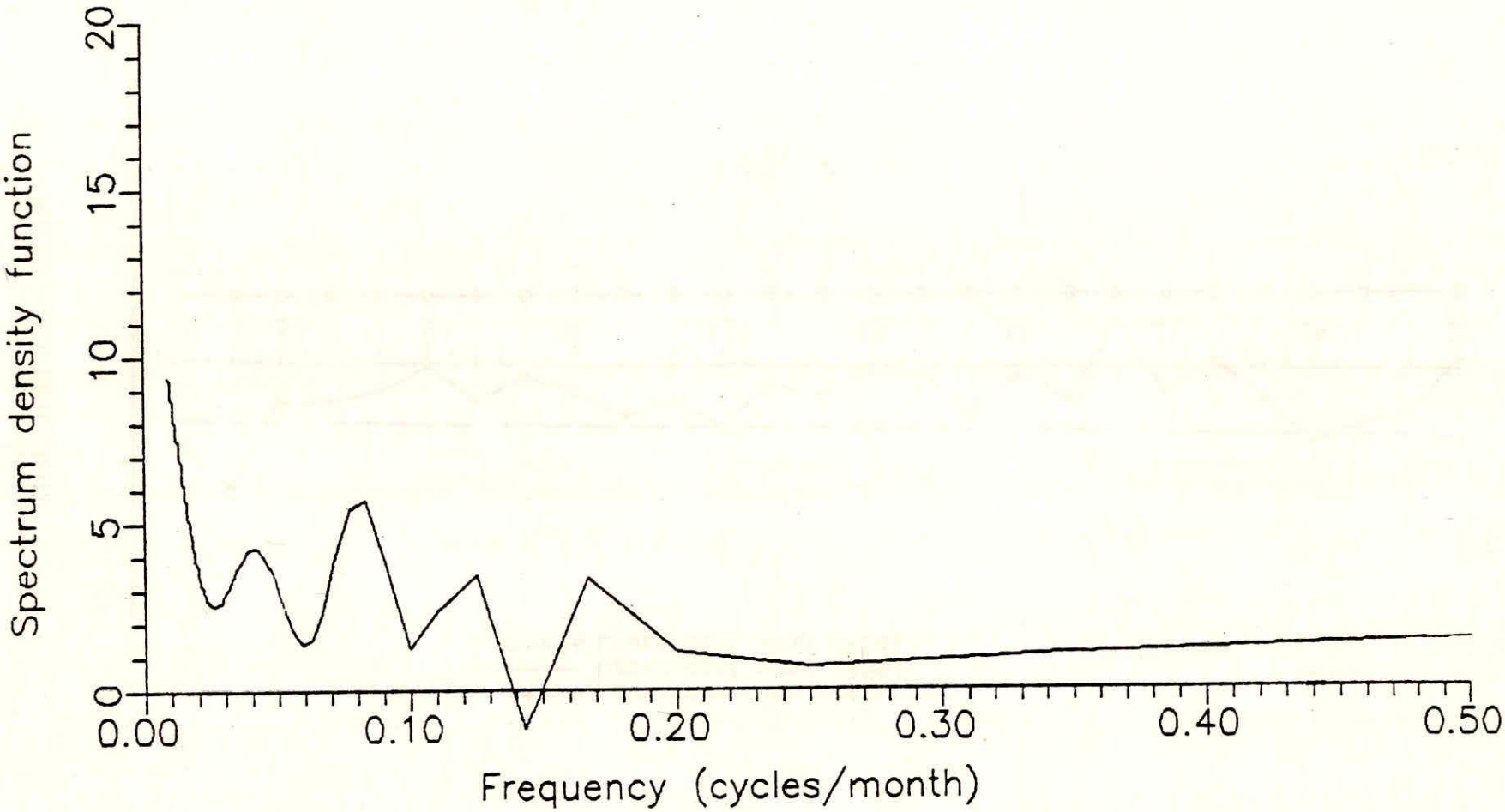

Fig. 7(a): Power spectrum of D/S DO series at Delhi on river Yamuna

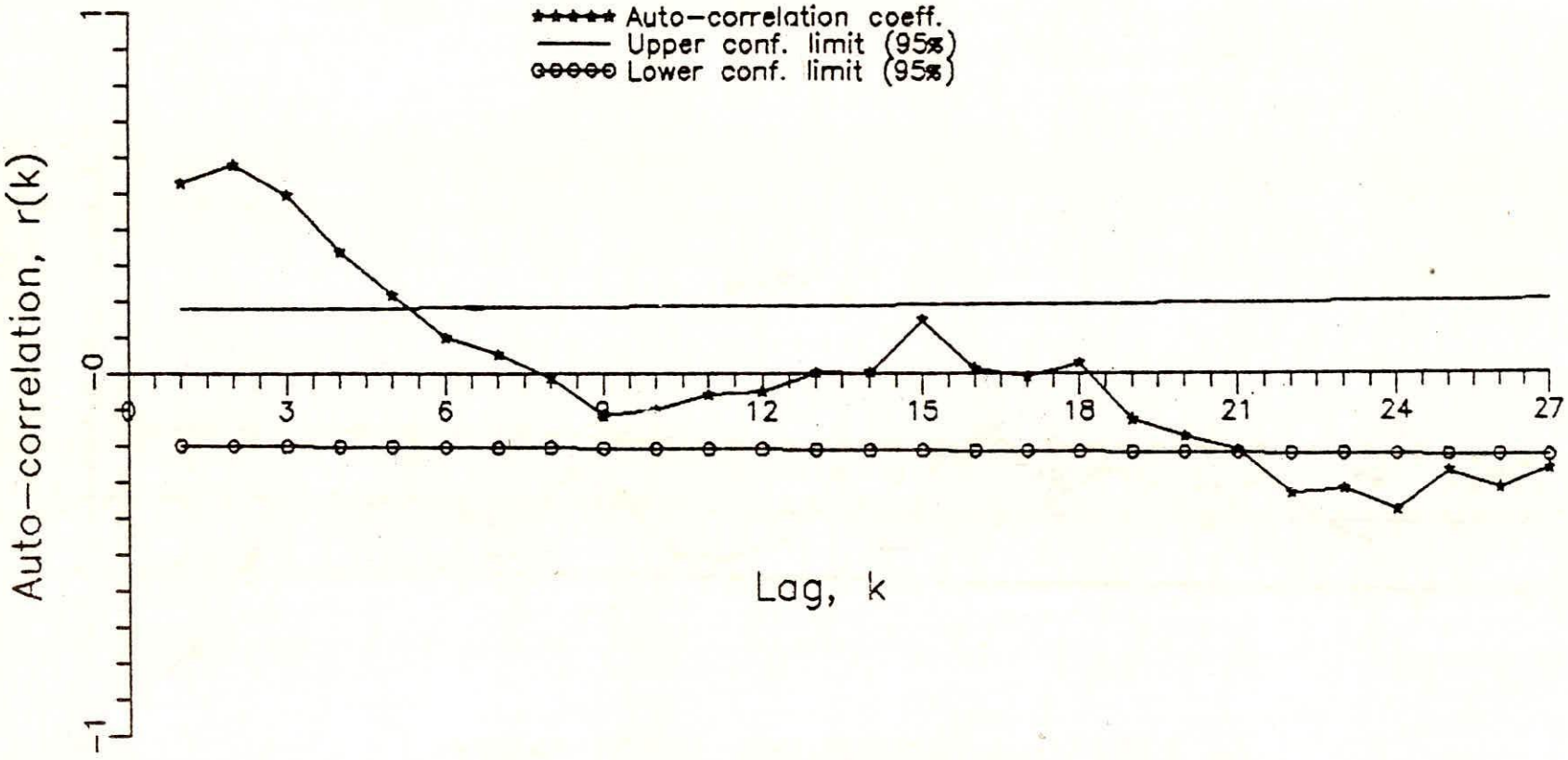

Fig.7(b): Auto-correlation function of DO series at U/S of Delhi on river Yamuna (after removing periodic component).

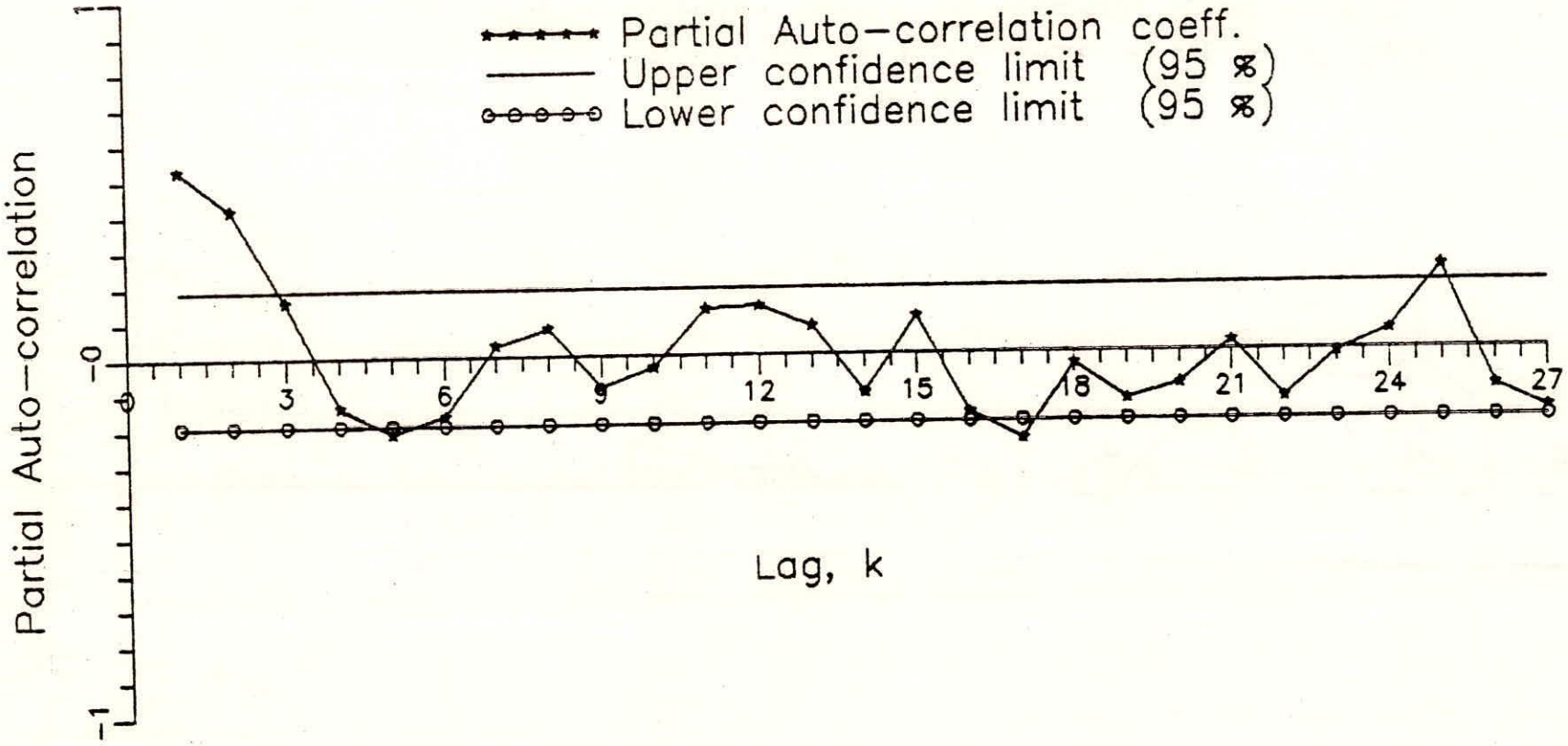

Fig. 8 : Partial Auto-correlation function of DO series at U/S of Delhi on river Yamuna (after removing periodic component).

Table 2. Summary of Harmonic Analysis of U/S Data

harmonic	A_i	B_i	P_i
1	-0.032	-0.464	0.017
2	-0.010	-0.016	0.017
3	0.301	0.330	0.032
4	-0.684	0.262	0.074
5	0.459	-0.265	0.096
6	-0.437	0.0	0.111

Similarly, for D/S section auto-correlogram (fig.6) and spectral density function (Fig.7.a) show the periodic characteristics of dissolved oxygen data. The Sp density function shows that all the six harmonics are significant and hence should be retained while substracting the periodic component from the original series. Table 3 summarises the results of harmonic analysis at D/S section.

Table 3: Summary of Harmonic analysis of D/S Data

harmonic	A_i	B_i	P_i
1	0.628	0.409	0.098
2	-0.024	-0.270	0.111
3	0.166	-0.075	0.117
4	-0.274	-0.019	0.130
5	0.047	0.530	0.180
6	0.008	0.00	0.180

4.3 **Dependent Stochastic Component:**

Several ARMA & ARIMA models were used to describe the dependent structure of the stochastic component from the trend & periodicity free series.

For U/S section, auto-correlation function (acf) and partial auto-correlation function (pacf)were calculated for lags 1 through 27.The acf and pacf were plotted in figs.7b and 8. The estimated acf and pacf suggested a higher order model which is prctically not justified as the dissolved oxygen depend on the past days rather than past months. Further, the estimated acf drops off slowly toward zero. which gives an indication of

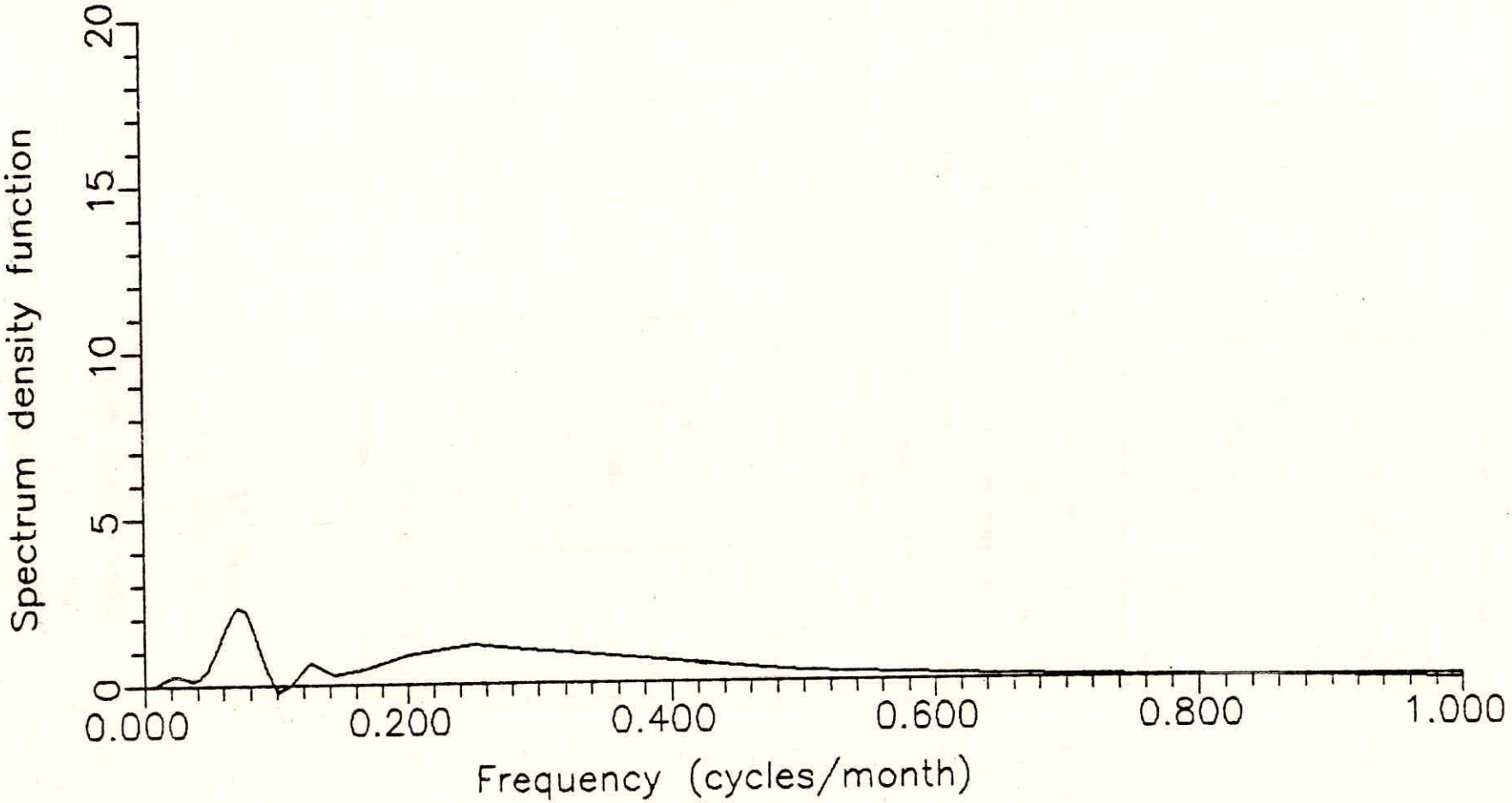

Fig. 9 : Power spectrum of first differences of DO data at U/S of Yamuna in Delhi.

Note : The statistical parameters of this figure are based on the first differences of the DO data.

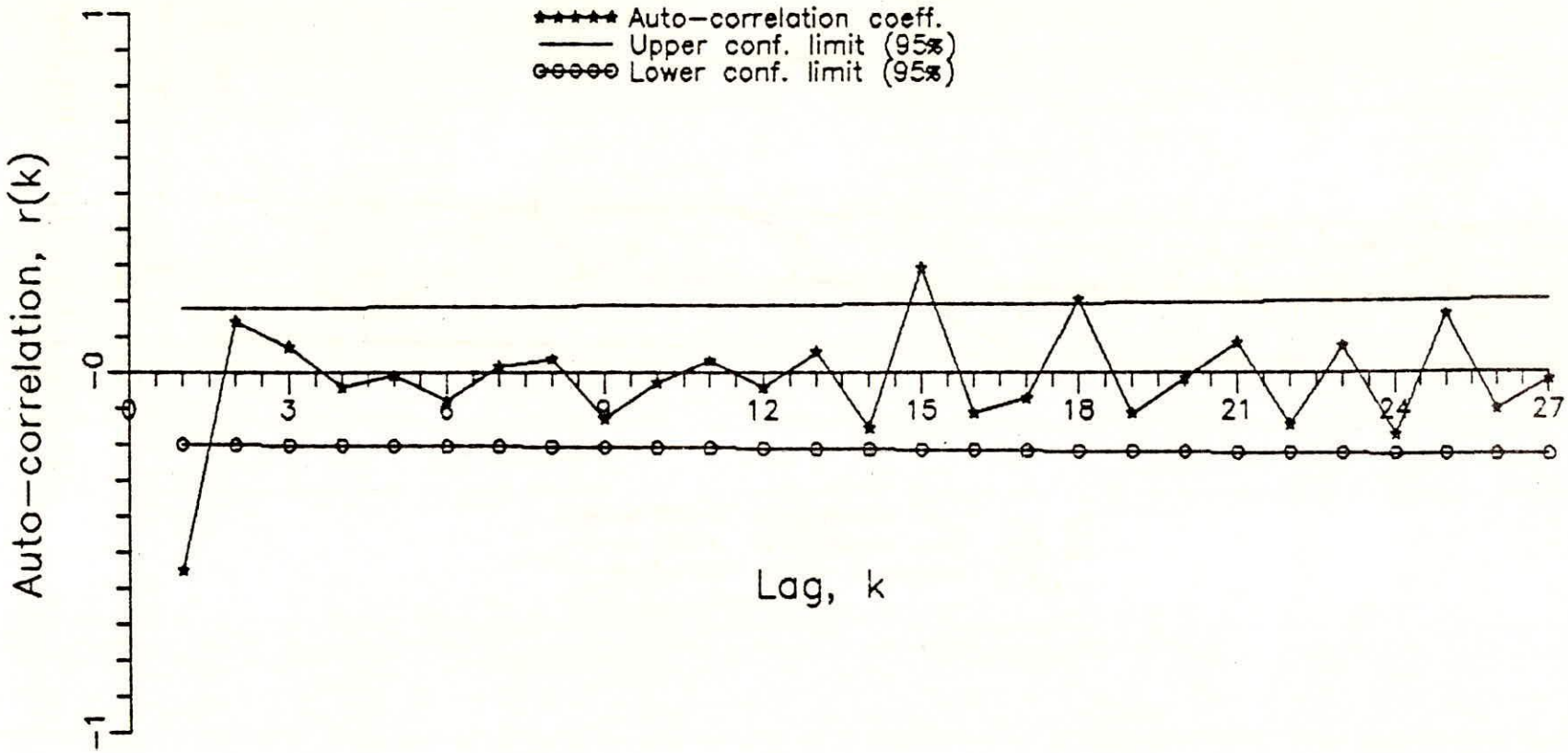

Fig.10 : Auto–correlation function of DO series at U/S of Delhi on river Yamuna (after removing periodic component).

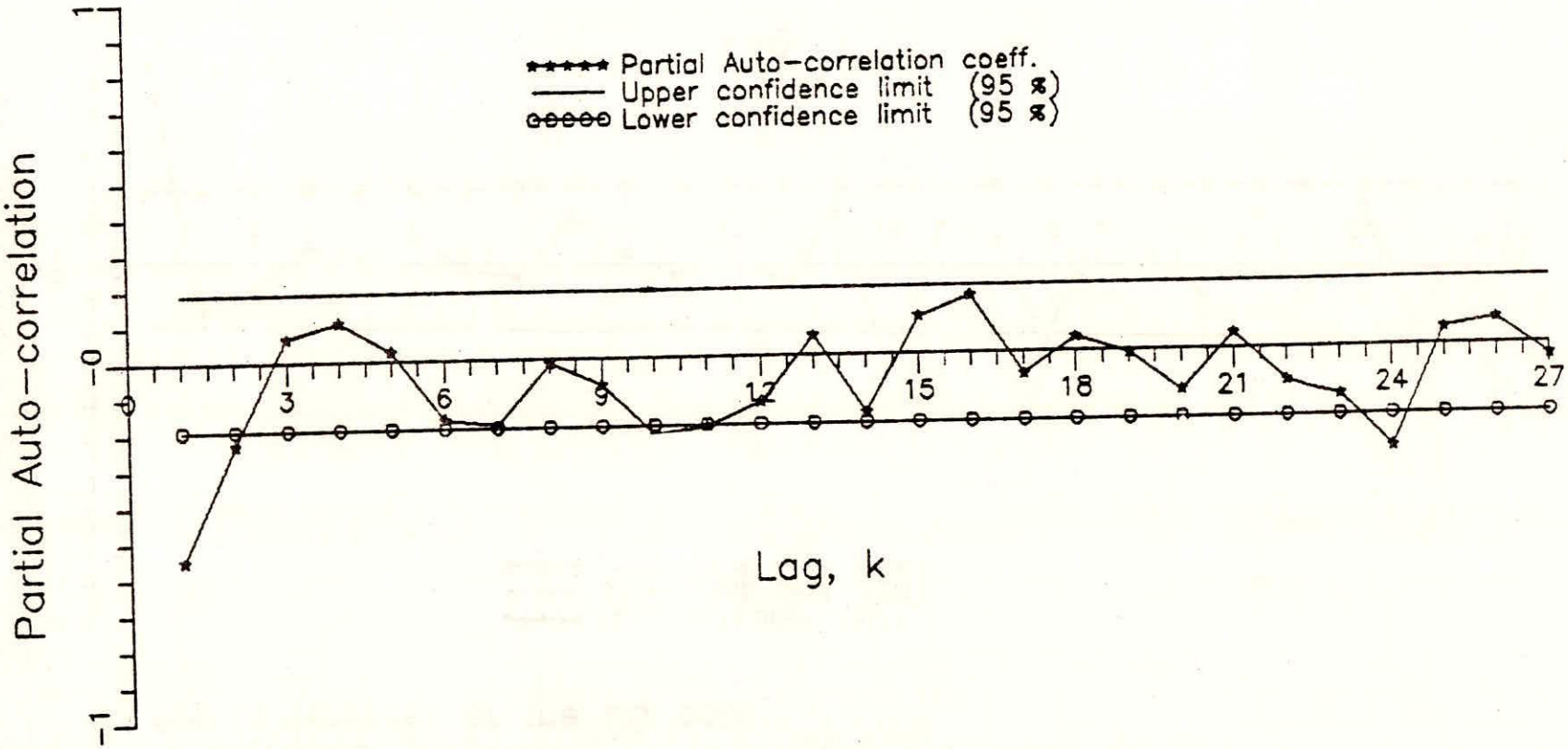

Fig. 11 : Partial Auto-correlation function of the first differences of DO data at U/S of Delhi on river Yamuna (after removing periodic component).

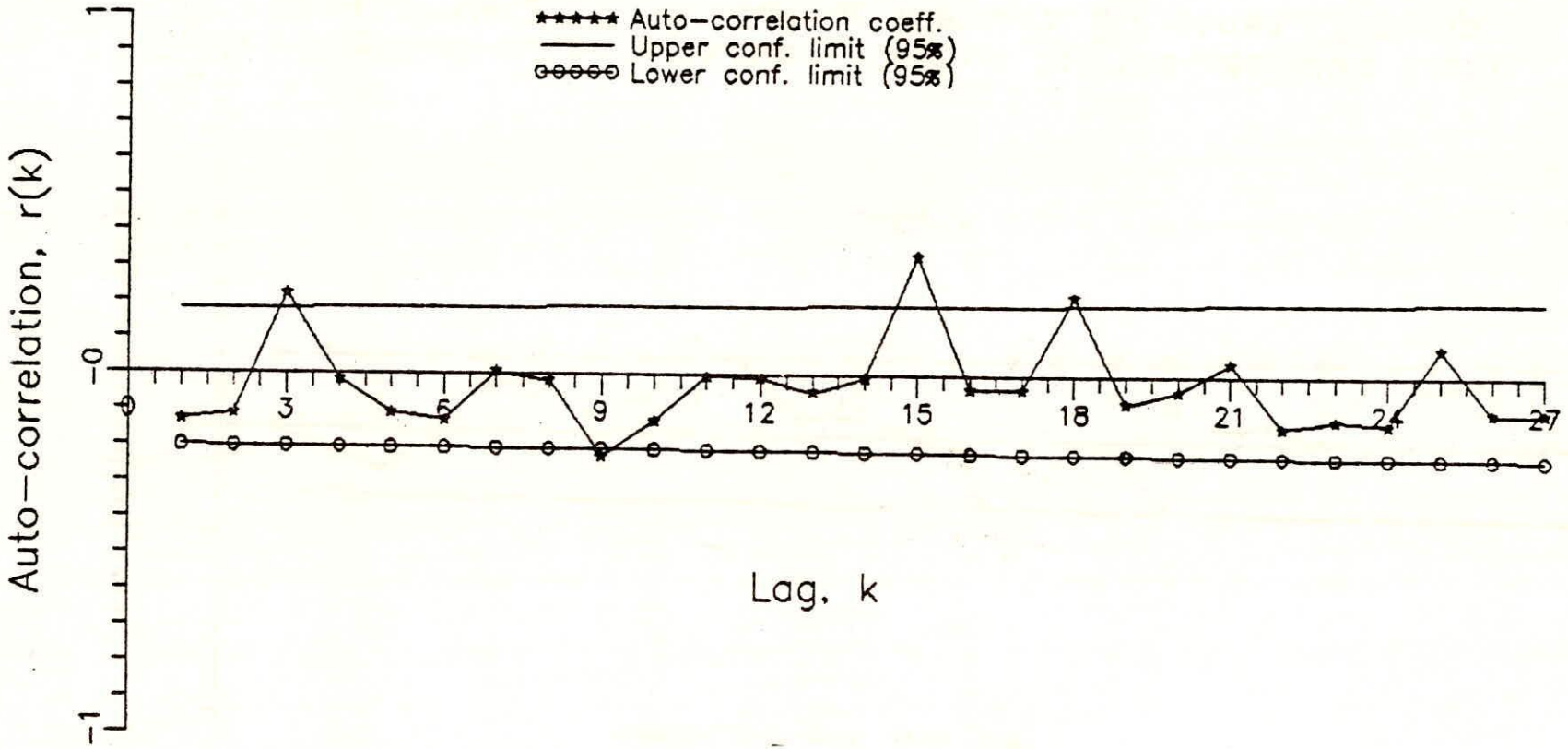

Fig. 12 : Residual auto-correlation function for the residuals from the ARIMA(1,1,0) model for the U/S DO series, Yamuna, Delhi

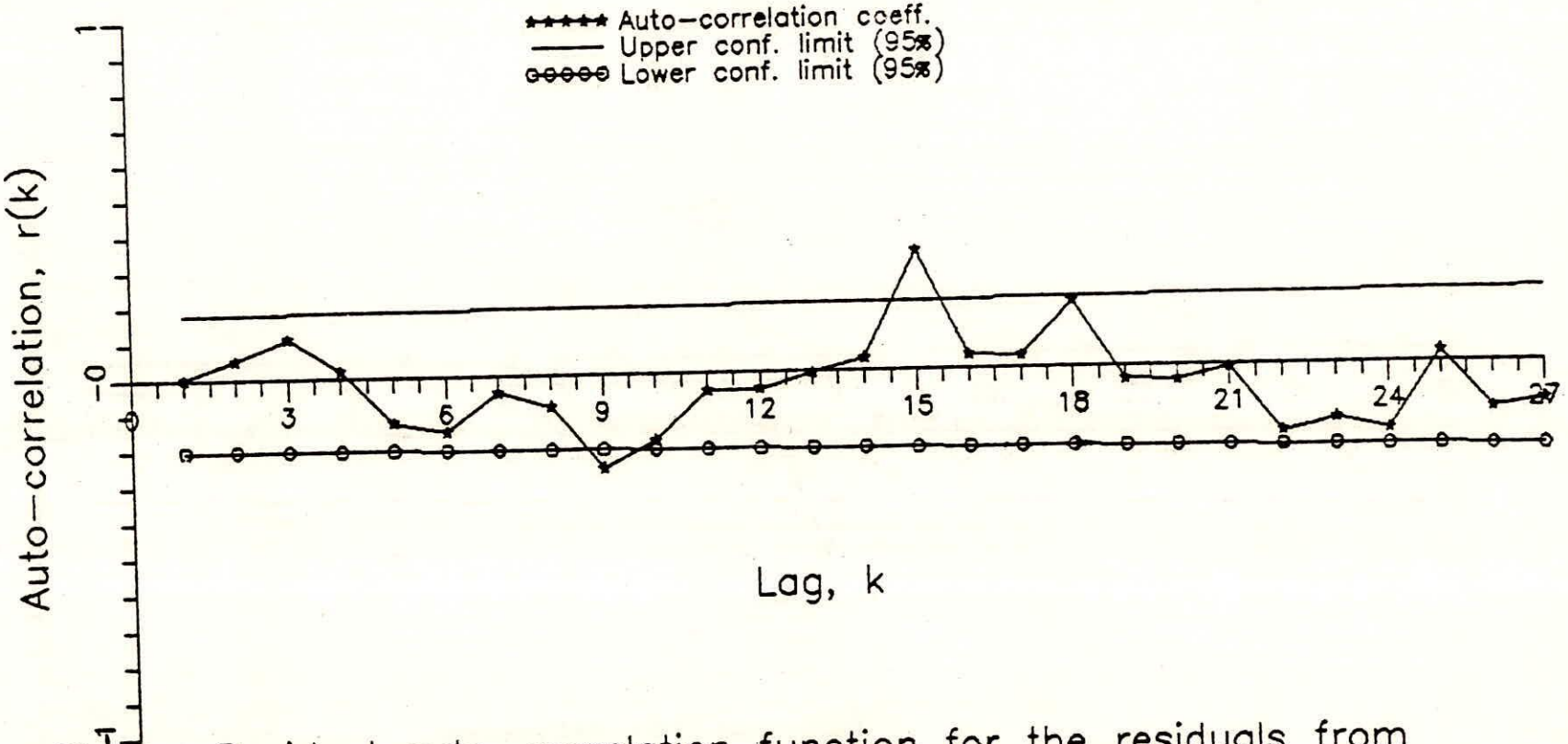

Fig. 13 : Residual auto-correlation function for the residuals from the ARIMA(2,1,0) model for the U/S DO series, Yamuna, Delhi.

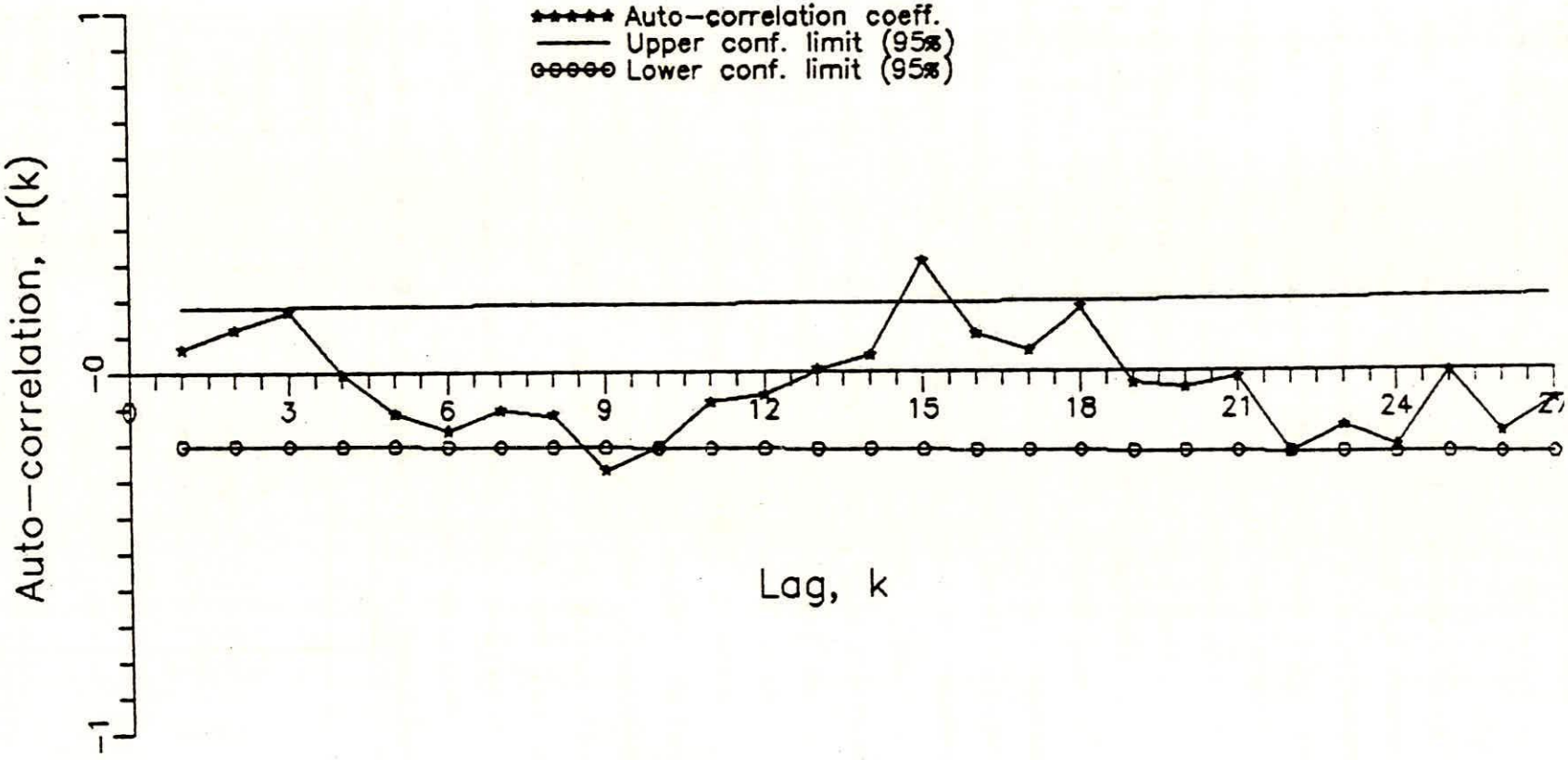

Fig.14 : Residual auto-correlation function for the residuals from the ARIMA(1,1,1) model for the U/S DO series, Yamuna, Delhi

non-stationarity character of series mean. The differencing technique were applied to remove non stationarity as suggested by Box-Jenkins(1976).

The first differences of DO data at U/S do not show any periodicity as shown by Sp.density function plotted in fig.9,which was removed automatically by differencing and hence no harmonic analysis was required for the differenced data. The estimated acf of first differences of DO data drops off rapidly to zero showing that the non stationarity is no more present in the differenced series. The acf and pacf were plotted in figs. 10 and 11 respectively. On the basis of estimated acf and pacf it appears reasonable to try the ARIMA(1,1,0), ARIMA(2,1,0),and ARIMA(1,1,1) models. The residual acf were used for diagonostic checking to test the hypothesis that the shocks of the applied model are statistically independent. The residual acf were calculated along with their t-values. The residual acf were plotted for ARIMA(1,1,0), ARIMA(2,1,0),and ARIMA(1,1,1) in figs. 12,13,and 14 respectively. Out of these three models ARIMA(2,1,0) is the best choice as the t-value for first 5 lags is insignificant in the residual acf of ARIMA(2,1,0),while ARIMA(1,1,0) residual has a significant t-value (more than 2) at lags 3,6,9,and 15 and ARIMA(1,1,1) residuals has t-value in the same pattern as ARIMA(2,1,0).Further, it is clear that among the ARIMA(2,1,0) and ARIMA(1,1,1) models, the ARIMA(2,1,0) is the best alternative as it has smaller t-values in the initial lags. The residuals of ARIMA(2,1,0) is tabulated in table 4.

Table 4: Residual acf of ARIMA(2,1,0) for U/S section

lag	acf	t-value
1	0.012	0.072
2	0.058	0.600
3	0.137	1.417
4	-.036	-.365
5	-.182	-1.84
6	-.226	-2.23
7	-.115	-1.08
8	-.112	-1.04
9	-.246	-2.27

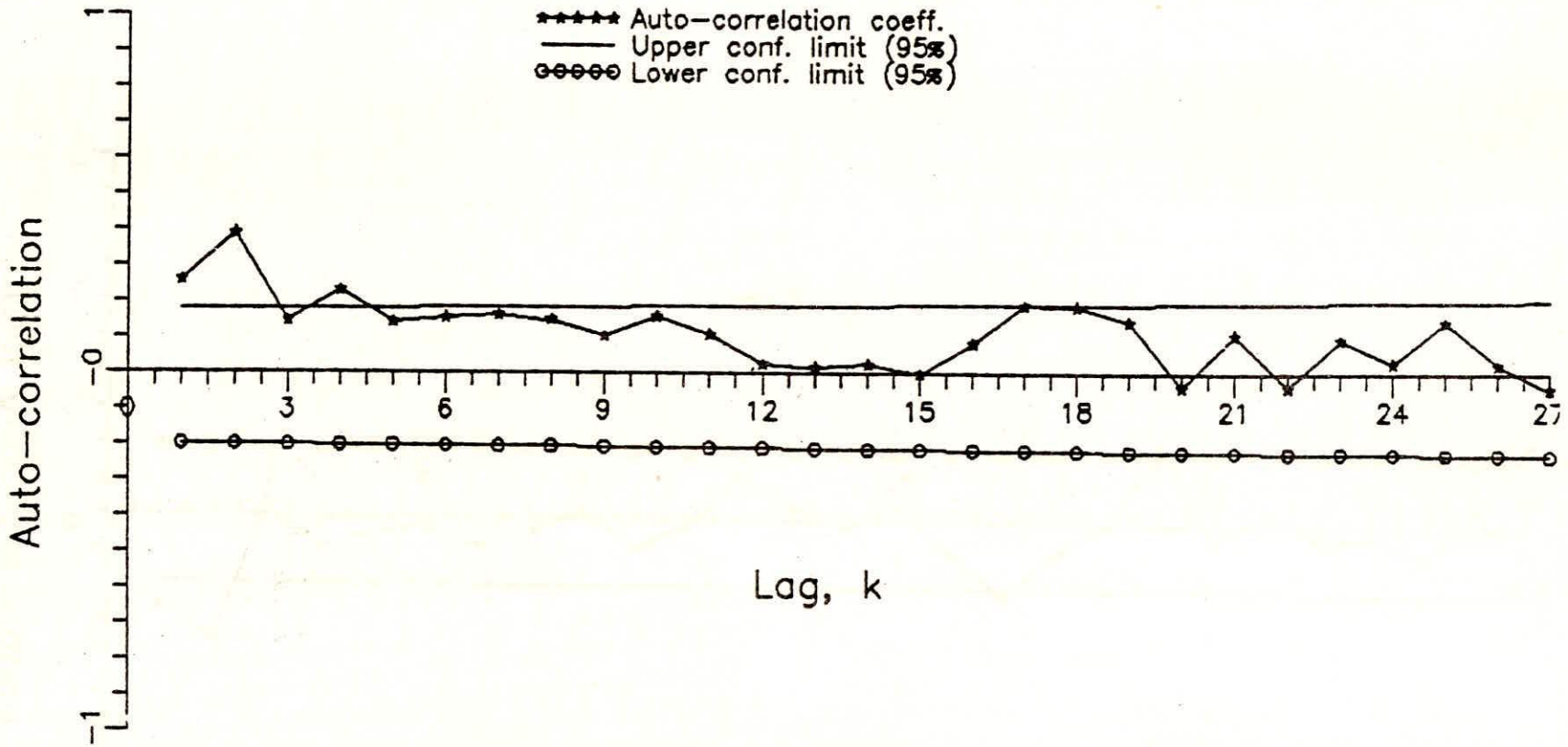

Fig. 15 : Auto–correlation function of DO series at D/S of Delhi on river Yamuna (after removing periodic component).

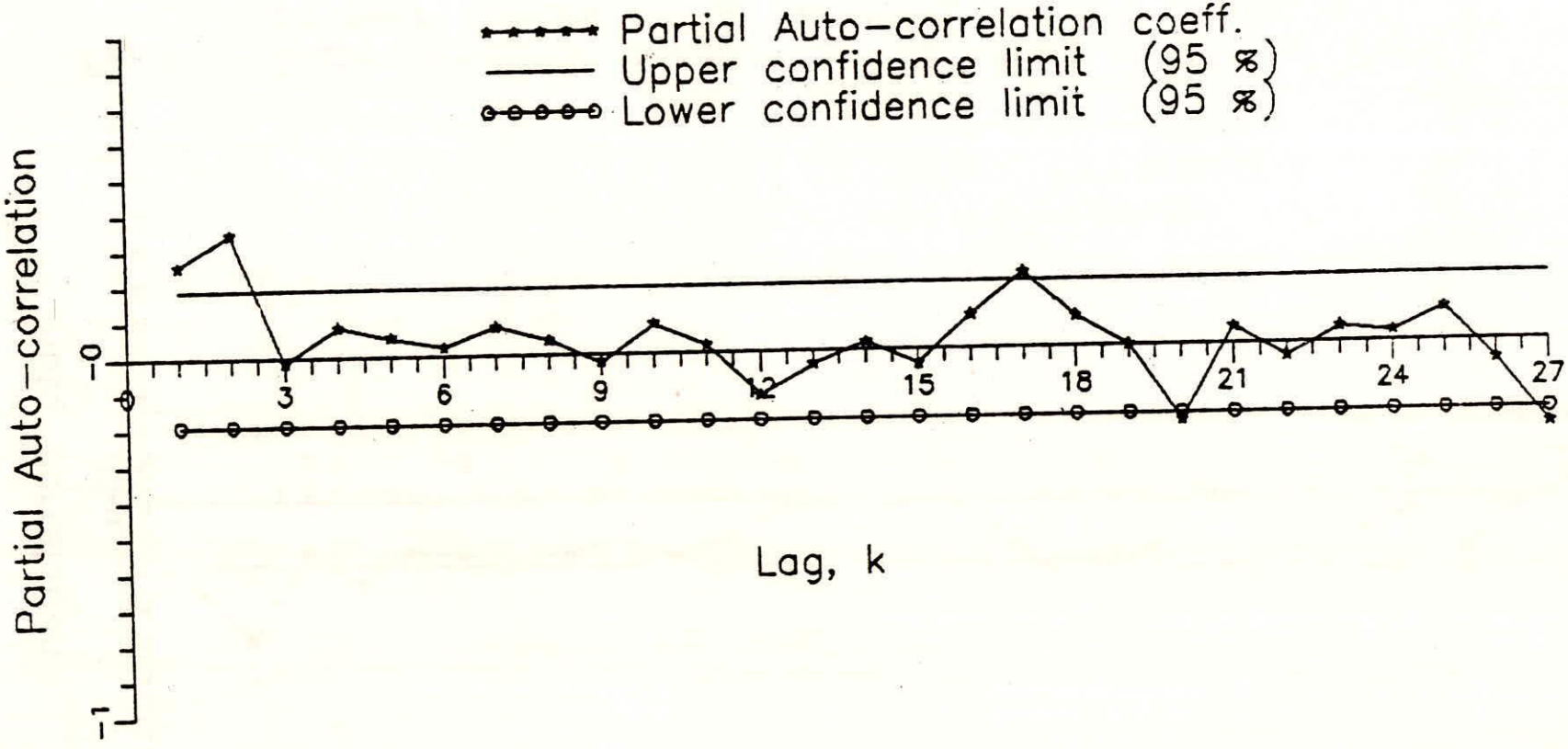

Fig.16 : Partial Auto-correlation function of DO series at D/S of Delhi on river Yamuna (after removing periodic component).

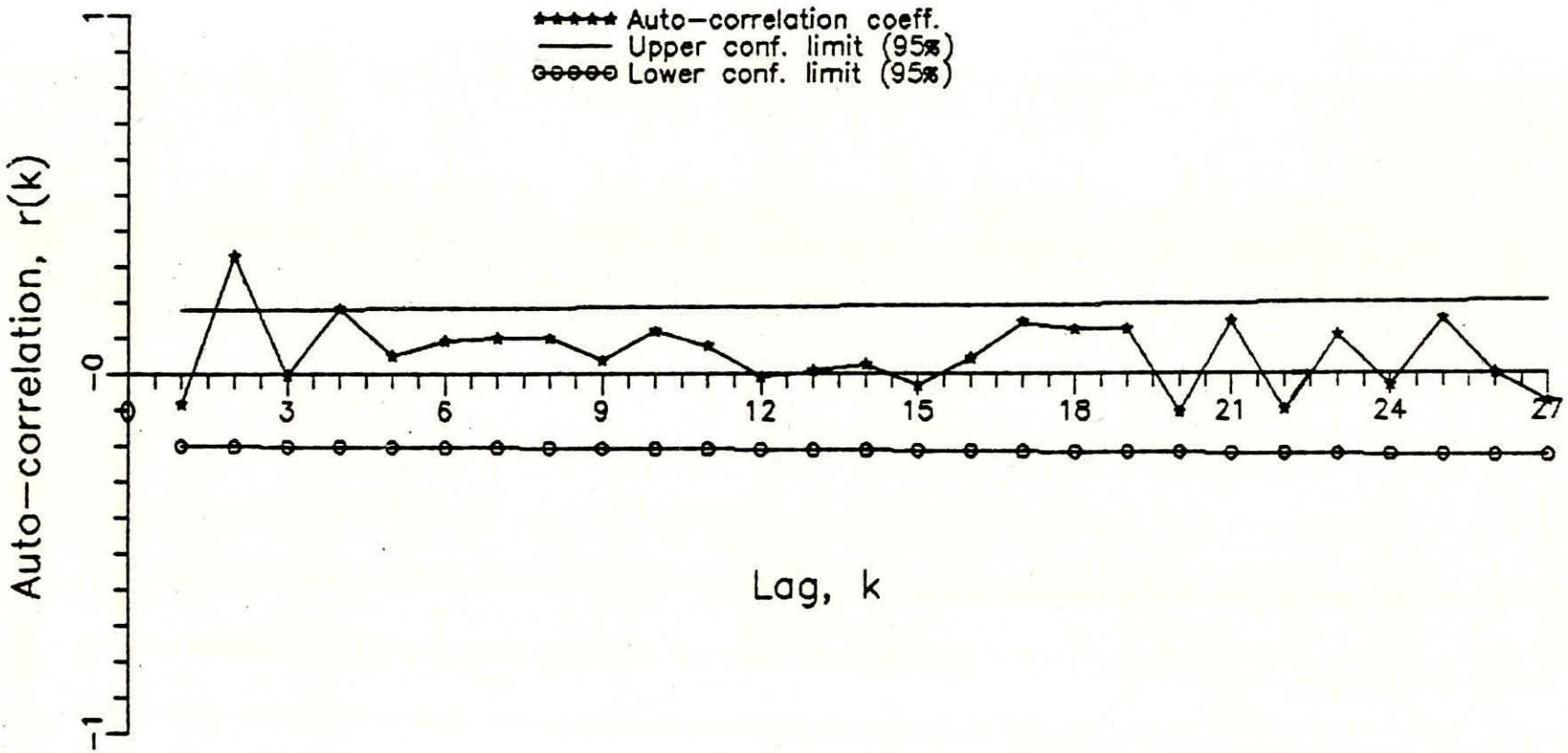

Fig.17 : Residual auto-correlation function for the residuals from the AR(1) model for the D/S DO series, Yamuna,Delhi.

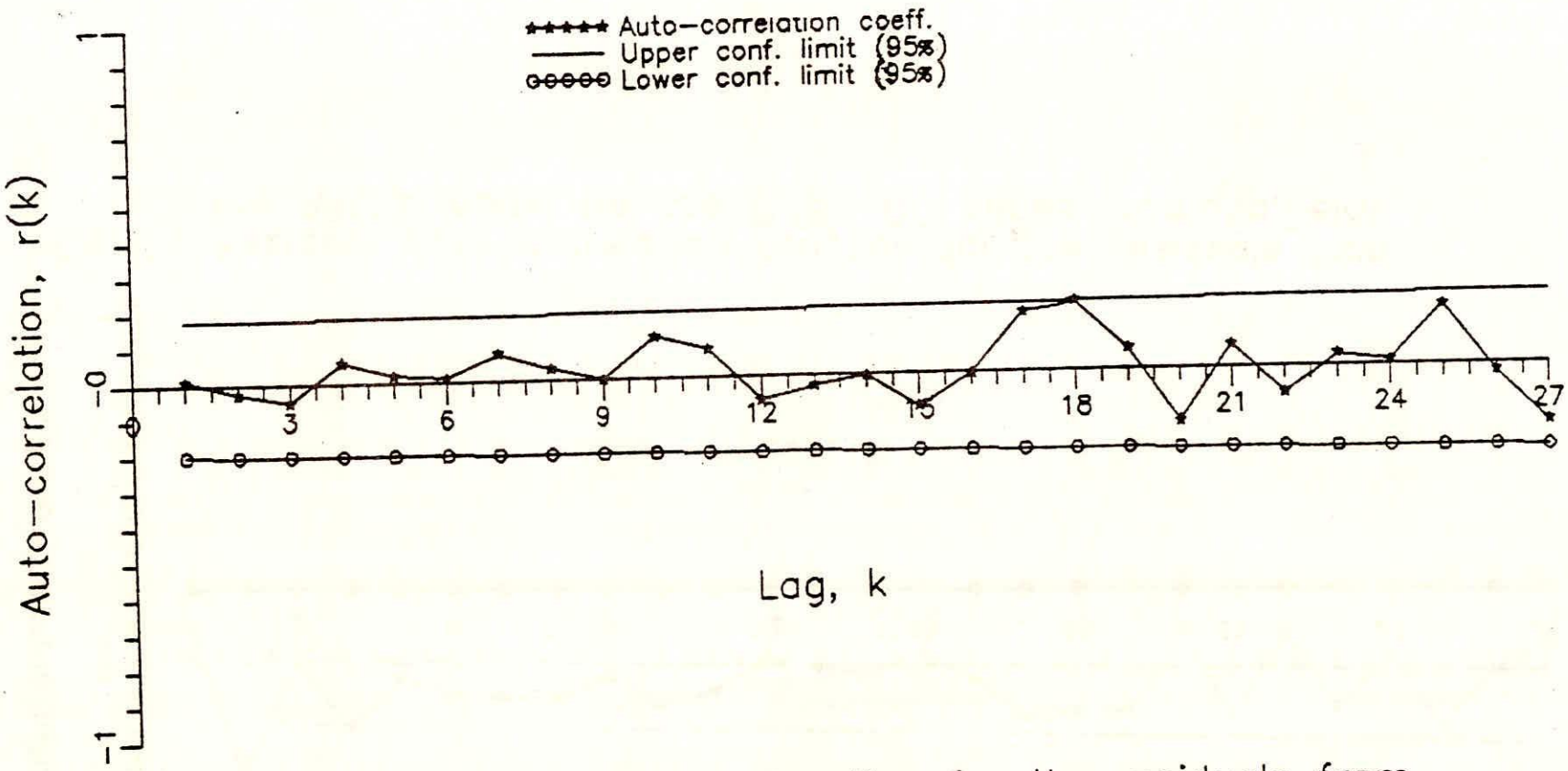

Fig.16 : Residual auto-correlation function for the residuals from the AR(2) model for the D/S DO series, Yamuna,Delhi.

10	-.114	-1.00
11	0.062	0.546
12	0.076	0.665
13	0.101	0.873
14	0.084	0.728
15	0.336	2.881
16	-.002	-.019
17	-.082	-.658
18	0.101	0.804
19	-.140	-1.107
20	-.142	-1.112
21	-.038	-.292
22	-.146	-1.124

The coefficients of ARIMA(2,1,0) model were estimated using the maximum likelyhood criterion. The estimated coefficients were:

$$\phi_1 = -0.588 \quad \text{and} \quad \phi_2 = -0.15$$

This model satisfy the stationarity requirements:

$$|\phi_2| < 1$$
$$|\phi_2 + \phi_1| < 1$$
$$|\phi_2 - \phi_1| < 1$$

For D/S section acf and pacf were calculated for lag 1 to 27 and plotted in figs 15 and 16. The estmated acf drops off quickly to zero which shows that the mean of the series is stationary and hence there is no need of differencing. Further, both acf and pacf coefficients are significant up to lag 2 and then oscillates with in the confidence band suggesting AR models to explain the D/S DO data.

AR(1) and AR(2) models were tried and the residual acf are shown in figs. 17 and 18. The estimated AR(1) coefficient ϕ_1 (equal to 0.25) is significantly different from zero and satisfy the stationarity condition. However, the residual acf is not good at all : the absolute t-value at lag 2 is 3.39 which is far larger than the residual acf short lag warning level of 1.25. This indicate that AR(1) model fail to give independent residuals and hence it is not an adequate model to explain the D/S data.

Fig. 18 and table 5 show that AR(2) model is satisfactory. All estmated coefficients ($\phi_1 =$ 0.167 and $\phi_2 =$ 0.334) are

statiscally significant and they meet the stationarity conditions. Diagonostic checking using the residual acf indicates that AR(2) is an adequate model:as none of the residual auto-correlations in table 5 and fig. 18 has an absolute t-value larger than our practical warning value(1.25). Furthermore, according to the chi-square test the residual auto-correlations are not significantly different from zero as a set. The estimated chi-squared statistics (equal to 22) is not significant. For 25 degree of freedom this statistic would have to exceed 34 to indicate statistical dependence in the random shocks at the 10 % level.

Table 5: Residual acf of ARIMA(2,0,0) for D/S section

lag	acf	T-value
1	0.014	0.084
2	-.023	-.232
3	-.050	-.519
4	0.061	0.631
5	0.022	0.222
6	0.014	0.148
7	0.078	0.793
8	0.034	0.348
9	0.001	0.008
10	0.116	1.180
11	0.080	0.797
12	-.066	-.660
13	-.029	0.292
14	-.005	-.053
15	-.093	-.917
16	-.003	-.034
17	0.167	1.640
18	0.195	1.870
19	0.057	0.528
20	-.148	-1.36
21	0.064	0.528
22	-.074	-.672
23	0.032	0.291
24	0.012	0.111

25	0.168	1.516
26	-.022	-.195
27	-.164	-1.45

4.4 Independent Residual Component:

The dependent stochastic part represneted by ARMA (p,q) model was substracted from the series. The remaining series containing independent stochastic part is called as the residual series which can only be described by some probability distribution function. In Box-Jenkins ARIMA modelling it is assumed that the independent residue component is normally distributed. Therefore, if the ARIMA modelling is correct the residual part must follow a normal distribution otherwise one should think to improve the ARIMA model so that the residual part may follow a normal distribution.

For U/S section the independent residue component has the following parameters:

Mean = 0.02

Stand. Dev. = 1.82

Skewness = 0.645

t-statistic = -1.98

For normality t-test was carried out, and the t-statistic is calculated by the following formula:

$$\text{t-statistic} = \frac{\text{calculated value} - \text{hypothised value}}{\text{standard error of estimate}}$$

where,estimated standard error is given by:

$$Se(g) = \sqrt{\frac{6N(N-1)}{(N-2)(N+1)(N+3)}} \qquad (53)$$

Using this equation, Se(g) for U/S = 0.2335 and t-value calculated using the equ.(53) is found to be -1.98 while critical t-value is 1.98 for 95 % confidence limit with two tailed test. Since calculated t-val is not smaller than critical t-value, we can't accept the hypothesis at 5%, level of significance, and hence the independent residue component can not be assumed normally distributed. As, already stated that the ARIMA model at the U/S section needs further improvment. However, as the calculated t-value for the sample is just equal to the critical t-value, one

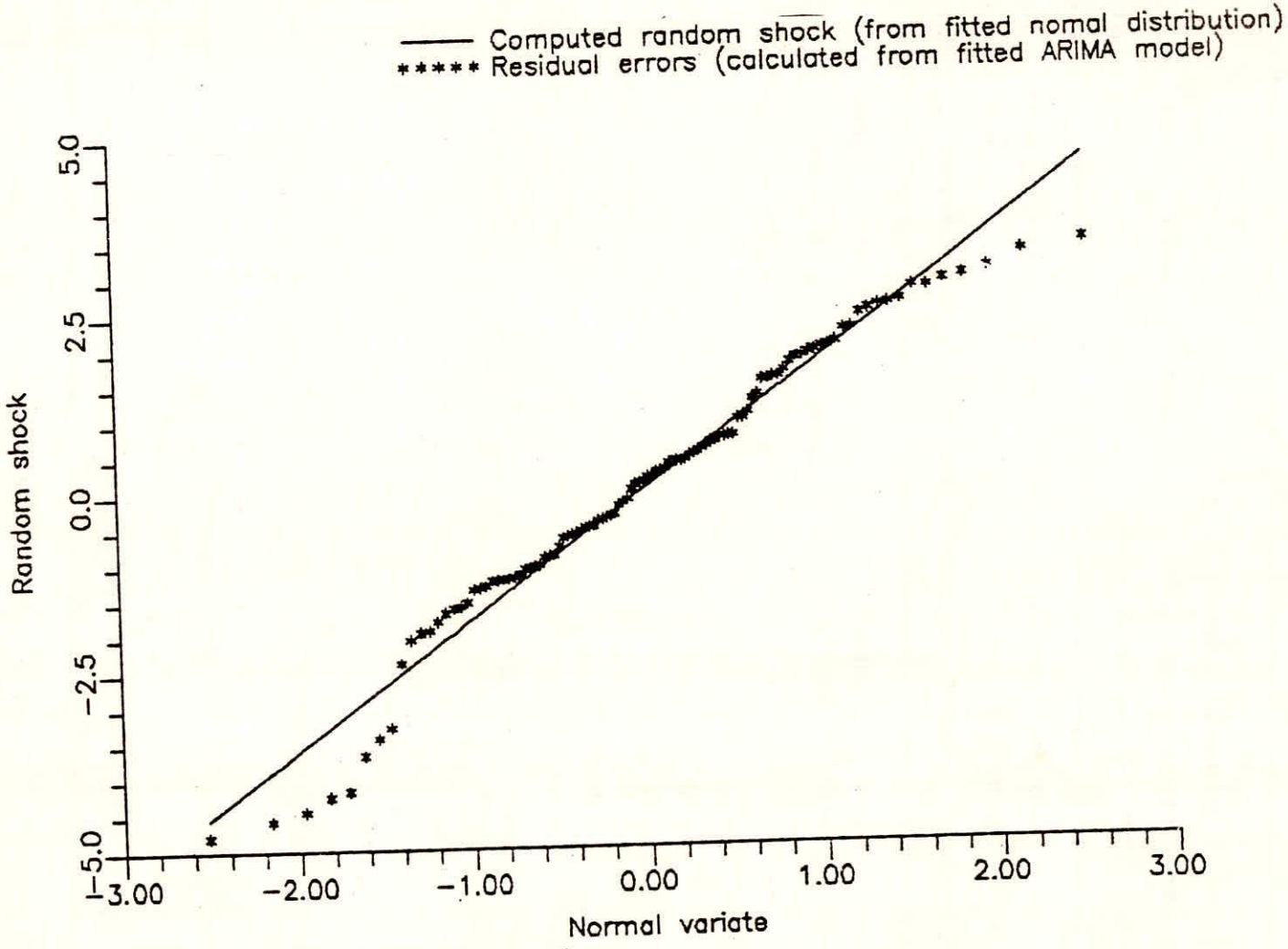

Fig.19 : Normal distribution fitted for the DO at U/S section of Yamuna at Delhi.

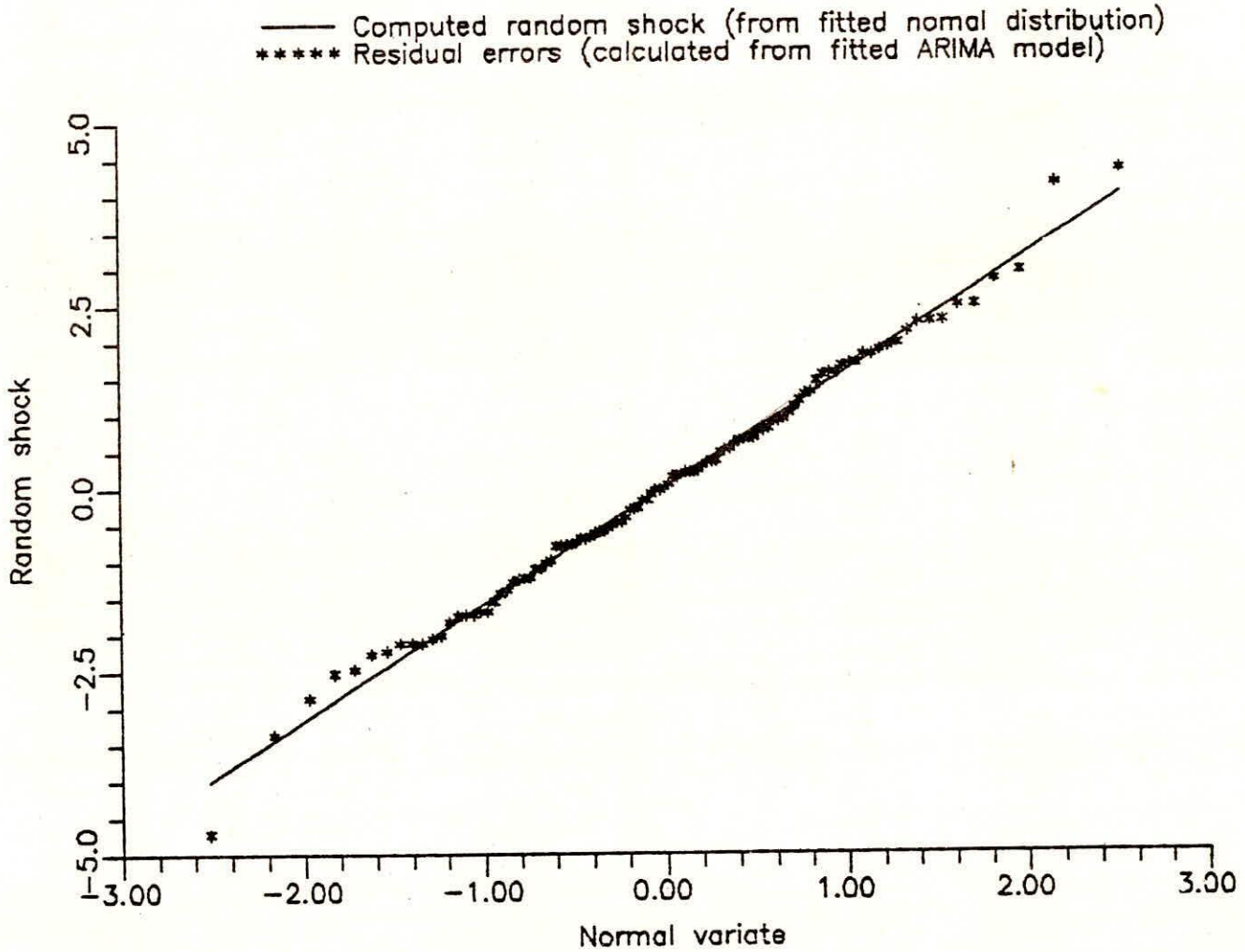

Fig.20 : Normal distribution fitted for the DO at D/S section of Yamuna at Delhi.

may roughly asume that the residue component is normally distributed. The independent residue can be generated using the following normal distribution equationevaluated by least square fitting:

$$a_t(c) = 0.0189 + 1.809\ z_{a_t} \qquad (54)$$

in which,

$$z_{a_t} = \text{reduced variate} = \frac{\hat{a}_t - \bar{a}_t}{\sigma_{a_t}}$$

$a_t(c)$ = independent residue

The coeff of co-relation is 0.997. The fitted normal distribution and the independent residual is shown in fig.19.

For D/s section the independent residual part has the following statistics

Mean	= 0.02
Stand. Dev.	= 1.58
Skewness	= 0.058
t-statistic	= 0.248

Because the calculated t-val is less than the critical t-value, the null hypothesis is accepted. The independent residue component can be generated using the following normal distribution:

$$a_t(c) = 0.0182 + 1.5895\ z_{a_t} \qquad (55)$$

where,

z_{a_t} = reduced variate

$a_t(c)$ = independent residue

The coeff. of co-relation is 0.999. The fitted normal distribution and independent residue is shown in fig. 20.

5.0 Results and discussion

In this study, an attempt has been made to analyze mean monthly dissolved oxygen concentrations observed at the U/S and D/S sections of the Yamuna river in Delhi. The results of the analysis are given below:

5.1 Deterministic component:

5.1.1 Trend component:

There was no indication of existance of trend in the annual data of dissolved oxygen concentrations at both the U/S and D/S sections. However, the monthly data at D/S section indicated the existance of trend. This may be due to the highly periodic nature of D/S DO data which was supposed to be removed in the analysis of annual data.

5.1.2 Periodicity:

The periodicity was detected in both the series at U/S and D/S sections with the help of auto-correlogram and spectral density function. At U/S section only first two harmonics were found significant while at D/S section all the six harmonics were found significant. This may be due to the reflection of seasonal charater of both the withdrawl of river water and waste disposal in between the U/S and D/S sections.

5.2 Stochastic component:

The stochastic component is comprised of dependent stochastic component and independent residue component. The results of both the component is summarised below.

5.2.1 Dependent stochastic component

The sesults of the chosen ARIMA models are given in the following table.

Table-6: Chosen ARIMA models and their parameters

station	ARIMA model	Parameters of model	Model
U/S	ARIMA(2,1,0)	$\phi_1 = -0.588$ $\phi_2 = -0.147, C=0.0$	$w_t = -.558w_{t-1} - .147w_{t-2} + a_t$ $w_t = \nabla^1 z_t = (1-B)^1 z_t$
D/S	ARIMA(2,0,0)	$\phi_1 = 0.167$ $\phi_2 = 0.334, C=0.0$	$z_t = .167z_{t-1} + .334z_{t-2} + a_t$

The ARIMA(2,1,0) model needs further improvements as it fail

to give residual auto-correlations that were significantly different from zero as a set (chi-square test). However, higher order model does not seem to be appropriate in the case of dissolved oxygen concentration as it depends on previous days but not on the previous 3 or 4 months. Further, there may be some another causes which influence the DO concentration at U/S section. For D/S section ARIMA(2,0,0) is an adequate model as it give residual auto-correlations that are not significantly different from zero as a set.

5.2.2 Independent residue component:

Independent residue component was represented by normal distributions as given in the following table:

Table-7: Charecteristics of residual series and parameters of chosen distributions

Characteristics of residual series obtained after ARIMA model					Parameter of fitted normal distribution $a_t(c)=\alpha+\beta z_{a_t}$			
obs. st.	mean	skewness		std. dev.	α	β	corr coeff	std.error of reg. equ.
		coeff	t-val					
U/S	0.02	-.4645	-1.98	1.82	.0139	1.809	0.997	0.30
D/S	0.02	0.0580	0.248	1.58	.0182	1.589	0.999	0.14

The residual series obtained after ARIMA model for U/S section are skewed towards right(skewness=-.4645) and calculated t-value is just equal to the critical t-value which indicate that the residuals have some what dependency which should be removed by further improving the ARIMA model. For D/S section, the residuals follow a normal distribution having more or less zero skewness (0.058) and insignificant t-value.

6.0 Conclusions

The following conclusions can be drawn from the work:

1. The D/S data series were adequately described by low-order ARMA models ($p,q \leq 2$). But for the U/S data, the residual series showed some dependence even after fitting ARIMA (2,1,0) model. A higher order model, ARIMA ($p,q > 2$), should therefore be used for this data series. The dependence of the residual series for low order ARIMA model for this series may imply that there are still some deterministic elements remaining in the series even after removing trend and periodicity by the methods described.

2. The two annual data sereis for U/S and D/S sections did not show any trend, which could be due to the reason of availability of small size of data (only 9 years), otherwise the increase in the effluent discharges from industrial, agricultural, and domestic sectors that have rapidly expanded in the recent past could have been reflected by showing the trend component. The seasonal variations in water quality is greatly influenced by the annual weather cycle and the cyclic pattern of the hydrological inputs of the river water environment. Uncertainties that are of a random nature, e.g., measurement errors, unexpected high levels of effluent discharges near the sampling site, and variation of the sampling point within the river cross-section, are reflected in the residue series.

3. The D/S series showed strong periodicity in comparison of U/S data series. This periodicity could be induced due to the periodic nature of raw water withdrawl and disposal of effluent discharges from various sources e.g. industrial, agricultural, and domestic sources etc.

4. The Box-Jenkins method for time series analysis was successful in modelling the monthly water quality data in the Yamuna river near Delhi. The models were parsimonious and physically reasonable. Dissolved oxygen data for U/S required a first difference-moving average model, while DO data for D/S section required moving average model without any difference.

5. The coefficients of selected ARIMA models were estimated by method of maximum likelyhood. However, for precise estimation of model coefficients one should use grid search technique c-

Marquardt's compromise scheme.

6. The Box-Jenkins technique was able to employ defective data, containing an oscillation believed to be machine-induced, to obtain a workable model. This method provides the water quality analyst with a new technique which may succeed where other methods would not. It can also serve as an alternative, and perhaps superior approach in situations where other methods can be employed.

APPENDIX - I

REFERENCES:

1. Streeter, H.W., and Phelps, E.B. (1925). "A Study of the pollution and National Purification of Ohio River." Public Heal .h Bulletin 146, USPHS, Washington, D.C.

2. Dobbins, W.E. (1964). "BOD and Oxygen Relationships in Streams." J.Sanit. Engrg. Div., ASCE, 90(3), 53-78.

3. Thayer, R.P., and Krutchkoff, R.G. (1967). "Stochastic Models for BOD and DO in Streams." J. Sanit. Engrg. Div., ASCE, 93(3), 59-72.

4. Thomann, R.V. (1967). "Time Series Analysis of Water Quality Data." J. Sanit. Engrg. Div., ASCE, 93(1), 1-23.

5. Box, G.E.P., and Jenkins, G.M. (1968). "Some Recent Advances in Forecasting and Control." Applied Statistics, Vol.17, No.2, 91-109.

6. Wastler, T.A., Walter, C.M. (1968). "Statistical Approach to Estuarine behaviour." J. Sanit. Engrg. Div., ASCE, 94(6), 1175-1194.

7. Dresnack, R., and Dobbins, W.E. (1968). "Numerical analysis of BOD and DO profiles." J. Sanit. Engrg. Div., ASCE, 94(5), 789-807.

8. Custer, S.W., and Krutchkoff, R.G. (1969). "Stochastic Model for BOD and DO in Estuaries." J. Sanit. Engrg. Div., 95(5), 865-885.

9. Kothandaraman, K., and B.B. Ewing (1969). "A Probabilistic Analysis of Dissolved Oxygen biochemical Oxygen Demand relationship in streams." J. Water Pollut. Control Fed., 41, R73-R90.

10. Kothandaraman, V. (1970). "Probabilistic variations in ultimate first stage BOD." J. Sanit Engrg. Div., ASCE, 96( ), 27-34.

11. Carlson, R.F., MacCormick, A.J.A., and Walts, D.G. (1970). "Application of Linear random models to four annual stream flow series." Water Resources Research, 6(4), 1070-1078.

12. Wallace, A.T., and Zollman, D.M. (1971). "Characterization of Time varying organic Loads." J. Sanit. Engrg. Div., ASCE, 97(3),

259-268.
13. Esen, I.I., and Bennet, J.P. (1971). "Probabilistic analysis of dissolved oxygen." Paper presented at the First International Symposium on Stochastic Hydraulics, Univ. of Pittsburgh, Pa.
14. Mc Mchael, F.C., and Hunter, J.S. (1972). "Stochastic modelling of Temperature and flow in rivers." Water Resources Research 8(1), 87-98.
15. Falkner, C.H. (1972). "DO prediction model for a long river." Water Resour. Res., 8(6), 1547-1559.
16. McMichael, F.E., and Vigani, F.E. (1972). discussion of "Characterization of Time Varying Organic loads." by A.T. Wallace and D.M. Zollman, J. Sanit. Engrg. Div., ASCE, 98(2), 444-455.
17. Soong, T.T. (1973). Random differential equations in Science and engineering. Academic Press, New York, NY.
18. McKerchar, A.I., and Delleur, J.W. (1974). "Application of seasonal parametric linear stochastic models to monthly flow data." Water Resour. Res., 10(2), 246-255.
19. Huck, P.M. and Farguhar, G.J. (1974). "Water Quality models using Box and Jenkins Method." J. Envir. Engrg. Div., 100(3), 733-753.
20. Shih, C.S. (1975). "Stochastic Water quality control by simulation." Water Resour. Bulletin, 11(3), 256-266.
21. Box, G.E.P., and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day, San-Francisco, California.
22. Padgett, W.J., G. Schultz, and C.P. Tsokos (1977). "A random differential equation approach to the probability distribution of BOD and DO in streams." SIAM, J. Appl. Math., 32, 467-483.
23. Padgett, W.J. and A.N.V. Rao (1979), "Estimation of BOD and DO probability distribution." J. Envir. Engrg. Div., ASCE, 105, 525-553.
24. Kottegoda, N.t. (1980). "Stochastic Water resources technology." John Wiley, New York, N Y.
25. Salas,J.D., Delleur,J.W., andYevjevich, V.(1980)."Applied Modelling of Hydrologic Time Series."Bookcrafters, Chelsea,Michigan,USA.
26. Finney, B.A., D.S. Bowels, and M.P. Windham (1982). "Random differential equation in river water quality modelling." Water

Resour. Res., 18, 122-134.

27. Pankratz, A.(1983)."forecasting with the Univariate Box-Jenkins Models."John wiley & sons, New York, USA.

28. Dewey, R.J. (1984). "Application of stochastic dissolved oxygen model." J.Envir. Engrg. Div., ASCE, 110, 412-429.

29. Gupta, R.K., and Chauhan, H.S., (1986). "Stochastic model of irrigation requirements." J.Irrig. and Drain, Engrg., ASCE, 112(1), 65-66.

30. Leduc, R.,Unny, T.E., and McBean, E.A. (1986). "Stochastic model of first order BOD kinetics." Water Res., 20(5), 625-632.

31. Lohani, B.N., and Wang, M.M. (1987). "Water quality data analysis in Chung Kang River." J. Envir. Engrg. Div., ASCE, 113(1), 186-195.

32. Zielinski, P.A. (1988). "Stochastic Dissolved Oxygen model." J.Envir. Engrg. Div., ASCE, 114(1), 74-90.

33. Jayawardena, A.W., Lai, F. (1989). "Time Series analysis of water quality data in Pearl River, China." J. Envir. Engrg. Div., ASCE, 115(3), 590-607.

34. Status report (1978-1990)."Water Quality Studies - Yamuna System." Central Water Commission,New Delhi-April 1991.

APPENDIX - II

NOTATIONS:

x_t	observation at time t
N	number of observations
r_k	auto-correlation coeff. at lag k
G(f)	spectral density function
f	frequency
M	maximum lag considered in the auto-correlogram
m_τ	computed periodic component of series
μ	population mean
h	total number of harmonic considered
A_i and B_i	Fourier coeff.for i^{th} harmonic
z_t	Time series after removal of Trend and periodic components
ϕ_{kk}	partial auto-correlation coeff. at lag k
ρ_k	hypothesized value of r_k
$S(r_k)$	estimated standard error for acf
$s(\phi_{kk})$	estimated standard error for pacf
ϕ_1, ϕ_2	AR(2) coefficients
θ_1, θ_2	MA(2) coefficients
a_t	random shock at time t
$r(\hat{a})$	residual acf
$S(r_k(\hat{a}))$	standard error of residual acf
Q^*	chi-squared test statistics
L	maximum likelihood function
Se(g)	estimated std error for normality test
$a_t(c)$	calculated independent residue at time 't'
z_{a_t}	reduced variate for random shock a_t
acf	auto-correlation function
pacf	partial auto-correlation function
AR	Auto-Regressive
MA	Moving Average
ARMA	Auto Regressive Moving Average
ARIMA	Auto regressive Intigrated Moving Avewrage

DIRECTOR : Dr.SATISH CHANDRA

DIV. SCIENTIST INCHARGE: SH.N.C.GHOSH

STUDY GROUP : ADITYA TYAGI