STOCHASTIC MODELLING IN HYDROLOGY

Any time series can be expressed as a linear combination of a trend component, a periodic component, a dependent 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

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.

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 = 1,2,3....,N, a turning point or p occurs at time t=i if x is either greater than x and x 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, p = the expected number of turning points in a random series = $\frac{2(N-2)}{3}$

Var(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 ,x ....., x. The standard procedure is to determine the number of times, p, in all pairs of observations (x ,x ;j > i) that x is greater than x ; 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 t defined as:

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

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

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

where, $\bar{\tau}$ = the expected number of t 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:

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

in which g 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 a and and its variance which are denoted by $\hat{\alpha}$ and $\hat{\sigma}_\alpha$ respectively; the statistic $t = \hat{\alpha}/\sigma_\alpha$ is then tested.

Periodicity analysis

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:

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

and

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

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

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:

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

where,

m_τ = the harmonically fitted means at period t(t =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 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 eq. 10 are $2\pi\tau/(p/i)$. The best estimate of the Fourier coefficients can be obtained by minimizing the $\Sigma(m_t - x_t)$, as given below:

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

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

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

For monthly data p=12, and therefore h=6. But for the most 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.

$$P = \frac{\sum_{i=1}^{j}[A^{2i} + B^{2i}]}{\frac{1}{p}\sum_{\tau=1}(x\tau - \mu)^2} \qquad (12)$$

where,

i = 1 to j, in decreasing order of magnitude

μ = the estimate of m is the mean of x

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

$$Z_t = X_t - m_\tau \qquad (13)$$

where,

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

m_t = periodic component of series X

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.l

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).

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.

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) is related to its own past values (z_{t-1}, z_{t-2}, z_{t-3}, ...). 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 $\varnothing_1$ z represent the predictable part of z while a is a residual element that cannot be predicted within the ARIMA model. However, the a 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 summarised schematically in Fig. 1.

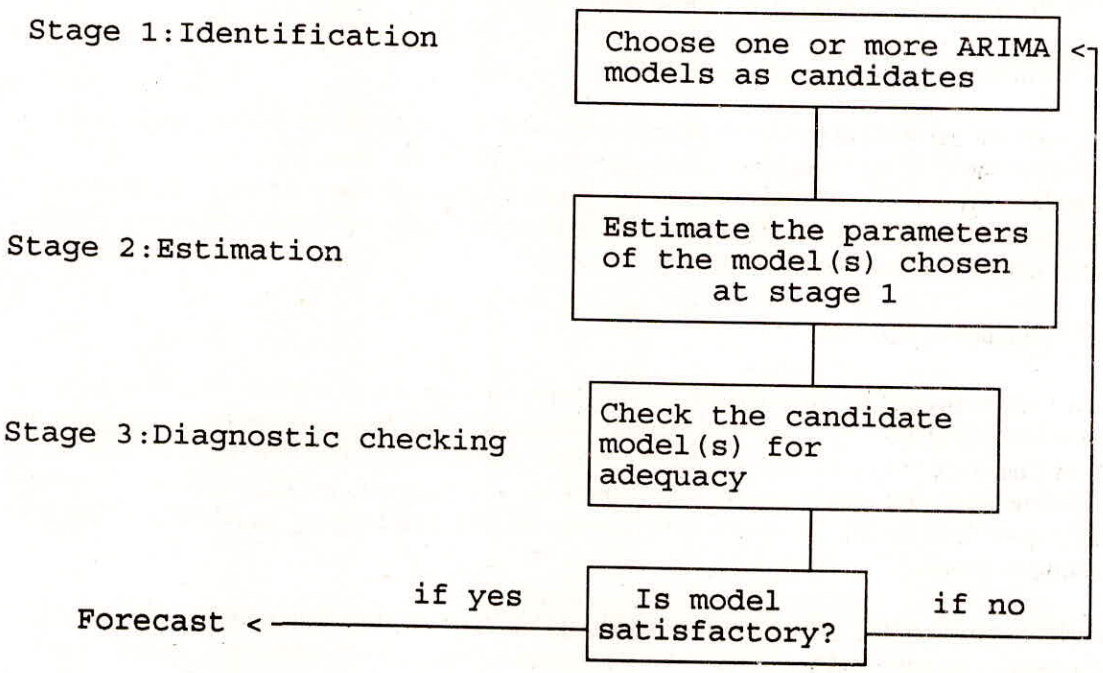

Fig.1

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 f 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.

Analytical tools for ARIMA Modelling

The two analytical tools estimated autocorrelation function (acf) and estimated partial auto-correlation 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 $(\bar{z}_t - \bar{z})$, 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 ordered pairs in question). The estimated partial autocorrelations coefficient measuring this relationship between Z_t and Z_{t+k} is desinged by 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 f coefficients. It involves using the previously calculated autocorrelation coeffecients(r).

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

of the partial autocorrelations.

$$\Phi_{11} = r \tag{16}$$

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

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

where,

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

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

For an independent series,the population correlogram is equal to zero for k=0. 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

$$r_k(99\%) = \frac{-1 \pm 2.29\sqrt{N-k-1}}{N-k} \tag{20}$$

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 significance. Any absolute t-value larger than 2 indicates that the corresponding r_k is significantly different from zero. The t-statistic for r_k is:

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

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

$$S(r_k) = \left(1+2\sum_{j=1}^{k-1} r_j^2\right) \quad (23)$$

the t-statistic for Φ_{kk} is:

$$t(\Phi_{kk}) = \frac{\hat{\Phi}_{kk} - \Phi_{kk}}{S(\Phi_{kk})} \quad (24)$$

where,

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

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

Modelling of Different ARIMA Models and Their Associated Charcteristics

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 in mind the principle 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 sinewave)	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 (i.e., 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:

$$z_t = c + \Phi_1 z_{t-1} + a_t \tag{26}$$

in backshift form this can be written as follows:

$$(1 - \Phi_1 B)\, \bar{z}_t = a_t \tag{27}$$

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

$$|\Phi_1| < 1 \tag{28}$$

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

$$z_1 = c + \Phi_1 z_{t-1} + \Phi_2 z_{t-2} + a_t \tag{29}$$

$$(1 - \Phi_1 B - \Phi_2 B^2)\, \bar{z}_t = a_t \tag{30}$$

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

$$|\Phi_2| < 1 \tag{31}$$

$$\Phi_2 + \Phi_1 < 1$$
$$\Phi_2 + \Phi_1 < 1$$

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:

$$z_t = c - \theta_1 a_{t-1} + a_t \tag{32}$$

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

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

$$(1 - \theta_t B)\, a_t = \bar{z}_t \tag{34}$$

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

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 ϕ_1 be less than one:

$$|\theta_1| < 1 \tag{36}$$

For MA(2) process the invertibility requirement is a set of conditions on ϕ_1 and ϕ_2:

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

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:

$$z = c + \phi_1 z_2 - \theta_1 a_{t-1} + a_t \tag{38}$$

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

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

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

$$(1 - \phi_1 B - \phi_2 B)\, \bar{z}_t = (1 - \theta_1 B - \theta_2 B)\, a_t \tag{41}$$

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.

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 minimum 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.

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:

$$r(\hat{a}) = \frac{\sum_{t=1}^{n-k} (\hat{a}_t - \bar{a})(\hat{a}_{t+k} - \bar{a})}{\sum_{t=1}^{n} (\hat{a}_t - \bar{a})^2} \qquad (42)$$

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 eq.(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, of course, 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-

$$H_o : \rho_i (a) = \rho_2 (a) = ... = \rho_K (a) = 0 \quad (45)$$

with the test statistic

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

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.

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).

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 about the origin is the product of its moment arm x and its mass y , and the sum of these moments over all the data is the

sample mean, given as follows-

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

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 $X = x_t$ 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:

$$\left[\prod_{i=1}^{N} f(x_i) \right] \qquad (48)$$

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 method 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.

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:

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

The properties of this distribution are-

i) that the variable $Y = \log_e y$ has a normal or ganssian distribution with mean m 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:

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

Some properties of this distribution are:

i) that its shape is determined by the two parameters a and p.
ii) that its mean is ap and its variance a^2 p.
iii) that it is unimodal for p > 1, skewed, with a 'tail' extending to the right.

The parameters a 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:

$$f(y)dy = \frac{1}{\sigma \Gamma p} \left(y - \frac{\alpha}{\sigma}\right)^{p-1} \exp\left[-\left(\frac{y-\alpha}{\sigma}\right)\right] dy, \alpha<y<\alpha, \alpha>0 \qquad (52)$$

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.

:: ::: ::