PAPER NO. 560

AN INTRODUCTION TIME SERIES MODELING AND ITS APPLICATION IN WATER RESOURCES ENGINEERING

BY

1. ENGR. TAHIR HAMEED 2. ENGR. TAHIR AHMAD MALIK
3. ENGR. MUHAMMAD SAFADAR CHISHTI

AN INTRODUCTION TO TIME SERIES MODELING AND ITS APPLICATION IN WATER ENGINEERING

BY

ENGR. TAHIR HAMEED(1) ,ENGR. TAHIR AHMAD MALIK(2) AND ENGR. MUHAMMAD SAFADAR CHISHTI(3)

ABSTRACT

Due to climatic and other uncertainties many time oriented processes in the field of water resources engineering cannot be represented by physical and chemical laws. However, time series theory provides an appropriate framework for adequate representation of such processes. An introduction of time series techniques with particular reference to autoregressive-integrated-moving-average (ARIMA) modelling and its potential application in the field of water resources are briefly discussed in this paper. For demonstration point of view, an ARIMA model is developed for a monthly discharge series of the River Indus observed at Kalabagh. The ARIMA model is found to be plausible and appropriate for such a series. The dependable forecasts from the ARIMA model can be used for a number of useful purposes.

INTRODUCTION

There are many variable in water resources engineering (such as precipitation and river flow) which are not governed by the known physical and chemical laws, but are dependent on the law of chance. Since such variables cannot be predicted in an exact sense, the water resources engineers has to be content with the probability laws. The statistical theory provides a reasonable framework for the modelling of such variables having inherent uncertainties.

A sequence of events that cannot be defined in an exact manner, is known as a stochastic process. A collection of data observed over successive intervals of time is called a time series. Such a data set consists of two group of numbers; the time at which each measurement is taken and the measurement itself. There are many examples of time series

(1) SRO/AEE, Irrigation Research Institute I&P Department Lahore.
(2) Chief Engineer, Irrigation Research Institute , I&P Department , Lahore.
(3) Assistant Professor, Uviversity of Engineering and Technology, Lahore.

data such as change of unemployment rate, share prices of a company, and population increase. In water resources some of the important time series are the discharge of a river observed at a particular location, total precipitation in a basin, water demand in an agricultural district or in a city.

A time series can be considered as a stochastic process and subsequently such processes can be studied with the help of the time series theory. The study of a series can yield some useful information. The past time oriented data of a series provides patterns which help to understand the causal mechanism of the process. Such indications at the identification stage assist in the development of a model that incorporate all of the patterns. Once a model is developed, it can be utilised for data generation and forecasting purposes. In the field of water resources engineering, the forecasts from a time series model may be utilised in planning and design of hydraulic systems and more importantly for operation of water resources.

There are advantages and disadvantages of time series modelling. Like all other data-based methods, the time series technique also has the assumption that the pattern which has been identified will continue in the future. If some major changes happen in future, the model developed based on past data will produce inaccurate predictions. Also, such models are empirical in nature and should strictly be applicable to similar data range and stations. But, unlike physical models which need observations of a number of independent variables and most of the time cannot forecast, only past observations of the variable of interest are sufficient in most cases for a time series model. The time series models produce fairly accurate forecasts which are not possible from a physical based model. However, more advanced and accurate time series models do need data of other variable apart from the variable of interest in order to establish relationships among themselves and for better forecast [3].

TIME SERIES THEORY

Classically, a time series is decomposed into two components as following:

$$X_t = p_t + E_t$$

where:

X_t = The observed time series,

p_t = The deterministic component,

E_t = Stationary stochastic component, and p can further be divided into three parts as:

$$p_t = t_t + o_t + s_t$$

in which

t_t = Trend or long-term movement,

o_t = Short-term oscillations or fluctuations, and

s_t = Periodic component or a component due to seasonality.

Auto-regressive-integrated-moving-average (ARIMA) models, popularised by Box and Jenkins [1], can effectively represent all of the above components. They suggested a step-wise procedure to develop ARIMA models that minimised the risk of errors which are more likely in classical time series modelling. The ARIMA models have already proven to be successful in the analysis of a number of time series in various fields.

The check for stationarity of the time series is the first step in the development of an ARIMA model. The weak form of a time series stationarity requires that the series has a constant mean, variance, and auto covariance function whereas for the case of the strong form of stationarity, the entire probability distribution function for the process has to be independent of time [5].

Such a form is not practical, and the weak form of stationarity provides a sufficient theoretical framework for a reasonably accurate analysis.

A non stationary time series can have a non-constant mean or a non-constant variance, or both of them. The term "non-constant" means that these statistics are dependent on time. A trend in a series produces a non-constant mean and such a series can be de-trended (and subsequently made stationary) simply by differencing. The differencing can be continued until the series produce a constant mean and variance, however, for most practical cases a maximum difference order of 2 is sufficient. For the case of seasonal data, the level of series shifts regularly up and down from the mean and these seasonal trends can be removed by seasonal differencing. The seasonal differencing depends on the nature of season over which the data are collected. For example, a difference order of 12 is needed for monthly data; 4 is for quarterly data and so on. The relationship between

observation at this lag (e.g., 12 for monthly data) are also included in the model to incorporate the seasonal patterns.

The non-constant variance can be stabilised by a suitable transformation. The Box-Cox transformation (Pankratz 1983) is a family of several transformations that is sufficient for most of the cases. The log and square root transformations are members of this family. A new series Z'_t is defined by the following relationship:

$$Z'_t = (Z_t^{\lambda} - 1) / \lambda$$

in which λ is a real number; and Z'_t is a non-stationary series.

The ARIMA model has auto-regressive (AR) and moving average (MA) parts. The middle letter, "I" in ARIMA, means integrated and represents the order of differencing for a non stationary series. If a difference order is zero, then such a series is already stationary and the resulting model is simply an ARIMA model. The AR part got the name auto-regressive because the variable $_t$ has a regression-type of relationship with its own previous values
Z_{t-1} , Z_{t-2} - ....., taken as independent variable as follows:

$$Z_t = \phi_1 Z_{t-1} - \phi_2 Z_{t-2} - .... - \phi_p Z_{t-p} + a_t$$

in which $\phi_1, \phi_2, ...,\phi_p$ = the first, second, .., and p-order AR coefficient, respectively; and
a_t = random shock error term or white noise, which is assumed to be normal and independently
distributed with a zero mean and a constant variance. The AR process has to satisfy certain stationarity conditions.

For example, for the first-order AR process [AR(1)], stationarity requires that the root of the process should lie outside of the unit circle or 1. For AR(2), the following conditions must be fulfilled:

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

In the moving average (MA) process, the variable of interest is explained through a random shock error term. The moving average model of order q [MA(q)] can be described as:

$$Z_t = -\theta_1 a_{t-1} - \theta_2 a_{t-2} - .... - \theta_q a_{t-q} + a_t$$

in which $\theta_1, \theta_2, ...\theta_q$= first, second,..., and q-order coefficients of the MA process, respectively. It is worthwhile to note that a_{t-q} is not exactly the past value of a_t but the random shock part of $_{t-q}$, which means that the MA process contains a part of the past value of $_t$ as well. Therefore, the MA process has an equivalent AR form and it can be proved that the MA process is equivalent to an infinite AR sequence [2]. Because of this characteristic, the stationarity conditions for an MA sequence are known as the invertibility conditions. For a MA(1) model, invertibility requires that $\theta_1 < 1$. The need for such a condition can be seen by a simple examination of the fact that a value of θ_1 more than 1 will assume that the recent observations have less influence on $_t$ than the further past ones. For a MA(2), model, the following invertibility conditions are required:

$|\theta_1| < 1$
$|\theta_2 + \theta_1| < 1$
$|\theta_2 - \theta_1| < 1$

The autocorrelation (ACF) and partial autocorrelation functions (PACF) show the measures of the strength of the relationship between observations at different lags. The ACF and PACF of an observed series help in the identification of an ARIMA model. The comparison of sample ACF and PACF with the commonly occurring theoretical ACFs and PAFCs leads to different possible choices. A variety of theoretical ACFs and PACFs are given in Box and Jenkins [1] and Pankratz (1983).

An ARIMA model is generally presented as ARIMA(p, d, q) in which p and q respectively represent the AR and the MA parts and d shows the difference order. The general ARIMA(p, d, q) can be described as:

$$\phi(B)(1-b)^d Z_t = C+\theta(B)\, a_t$$

in which C = a constant; B = the difference operator defined as $B^iZ_t = Z_{t-1}$. d = the difference order; $\phi(B) = 1-\phi_1B -\phi_2B^2 -,..... ,\phi_qB^q$ = the p-order AR operator; and $\theta(B) = 1-\theta_1B -\theta_2B^2 -,.... , \theta_qB^q$ = the q-order MA operator. The stationarity condition for ARIMA models requires that the roots of the characteristic equation of $\theta(B)=0$ lie outside of the unit circle. Similarly for invertibility, the roots of the characteristic equation $\theta(B)=0$ must lie outside the unit circle [4]

As already mentioned, in the case of seasonal data where the series may have seasonal relationships at lag s, simple ARIMA models may be inappropriate and inadequate. The lag s depends upon the nature of data. For example , in a monthly observed series, the relationship at a lag of 12 is important and may not be neglected for the development of an adequate model. Such relationships should be represented by seasonal ARIMA models which simultaneously include both seasonal and non-seasonal parts. The multiplicative type of seasonal models are proved to be sufficient for most of series. The standard notation for such models is ARIMA(p, d, q)(P, D, Q)s and can be represented as:

$$\phi(B)\Phi(B^S)(1-B)^d(1-B)s^D Z_t = C + \theta(B)\Theta(B^S)\, a_t$$

in which $\Phi(B^S) = 1 -\Phi_s(B^S) - \Phi_{2s}(B^{2S}) - ,\cdots\cdots\cdots,\Phi_{ps} B^{pS}$ = the P-order seasonal operator and $\Theta(B^S) = 1 -\Theta_s(B^S) - \Theta_{2s}(B^{2S}) - ,\cdots\cdots\cdots,\Theta_{ps} B^{pS}$ the Q-order seasonal MA operator. The stationary and invertibility conditions require that all the roots of the characteristic equations

$\Phi(B) =0$, $\Phi(B^S) = 0$, $\theta(B) = 0$, $\Theta(B^S)$ lie outside of the unit circle in the complete plane. Details regarding roots of characteris equation can be found in Box and Jenkins [1].

Box and Jenkins [1] recommended a three-stage strategy for the development of ARIMA models. In the first stage, identification of the correlation structure within a stationary series is done with the help of graphical representations of sample ACF and PACF. A variety of possible models, selected in the first stage, are then estimated in the second stage. The Akaike Information criteria and residual variances of different models can be used for the tentative selection of a particular model [5]. The most appropriate model is then verified in the diagnostic-checking third stage. If the selected model proves to be adequate and passes the various residual tests, then it may be used for forecasting

purposes.

Parsimony is another important consideration in the development of ARIMA models. The higher order AR or MA models can be represented by ARIMA models of fewer parameters. Box and Jenkins (1976) defined a good model as one having the fewest possible parameters.

APPLICATION IN WATER RESOURCES

As already mentioned in the introduction section, a number of time series can be found in the area of water resources. Some of the important time series in this field may be river flows observed at a location, municipal and agricultural water demand, sediment rate in a channel, and precipitation in a basin. Monthly discharge data of river Indus from January, 1975 to December, 1981, recorded at Kalabagh are used for an example of the development of an ARIMA model. The first 60 months (January, 1975 to December, 1980: shown in Fig. 1) are used for the model development and the last year data (January-December 1981: 12 months) are utilized for the comparison with the forecast from the final ARIMA model.

The series Q drawn in Fig. 1 shows a cyclical behavior where cycle repeats at a lag of 12 months, which is typical for the monthly observed data. It is clear from this figure that the variance is stable and the transformation of the original data is not needed.

The ACF and PACF of Q series are plotted in Figs. 2 and 3. The ACF for the series shows significant cyclical peaks at lags 12, 24, and 36. These peaks do not die out quickly meaning that the series has a seasonal nonstationarity which is the usual case for monthly data. The significant autocorrelations at half seasonal lags 6 and 8 can be due to strong seasonal variations and they often become insignificant after seasonal differencing and/or when seasonal AR and MA coefficients are estimated (pankratz 1983).

The seasonally differenced Q series (s = 12) is shown in Fig. 4. It is clear that the predominant original cyclical behavior is gone and the remaining part then can be modeled as ARIMA. Fig. 5 and 6 show the ACF and PACF of seasonally differenced Q series. The significant autocorrelations at lags 6 and 12 have disappeared and only big spikes at lags 12 and 24 are left. Such a pattern of big spikes is an indication of the need

for a seasonal MA term.

The remaining pattern shows a decaying trend which is the indication for the inclusion of a non seasonal AR term Normally a second-order auto-regressive (AR(2)) and some times a single-order auto-regressive[AR(1)] process can be adequate for such patterns. It is always a good practice to develop a model with minimum number of parameters. Keeping in view all of the above mentioned indications, following ARIMA model can be considered as possible choice:

ARIMA (1,0,0) (0,1, 1) 12

or

$$(1 - \phi_1 B)(Z_t - Z_{t-12}) = a_t - \theta_{12} a_{t-12}$$

The Model is then estimated by ASTSA computer package (Shumway 1988) as following:

$$(1 - 0.29\ B)(Z_t - Z_{t-12}) = a_t - 0.65\ a_{t-12}$$
$$(0.14) \qquad (0.12)$$
$$(2.13) \qquad (5.42)$$

Residual Variance = 4.97
AIC = 2.74

where the figure in the perenthesis directly below the model equation are respectively the standard error and t-value of the coefficient.

Various adequacy and residual test are then applied for the selected model and it is found to be clear from any abnormalities. Although this model can be improved by an in-depth analysis, however, it is considered to be adequate for demonstration purposes.

The forecasting capability of an ARIMA model is the ultimate criterion for its selection and capability. The last 12 observations (Jan-Dec. 1981), which were left for model calibration, are then compared with the forecasted values form the selected model. These

forecasts are obtained by one-step-ahead forecasting proccdure. In this procedure, the first forecast is started from a reference point (in the present study the reference point is December, 1980 and the first forecast is January, 1981) and the second forecast is obtained from the first forecast. the procedure continues until the required number of forecast are attained. Such a procedure yield the so called "bootstrap" forecast. This procedure allows the incorporation of the more recently observed values in the model, update the model parameters, and start the forecast from the newly observed last data point. Such additions offer two advantages: first, the model parameters are updated, and second, the future forecasts are better than those obtained without new additions. Also, because ARIMA models are better suited for short-term forecast (Box and Jenkins 1976); the incorporation of recently observed values strengthens the reliability of the forecasts.

Table 1 shows a comparison of the observed and forecasts values form the final selected model. The forecast error is the defference between observed and forecast values for the respective months. The forecast correctly track the seasonal psttern of series. The root mean squared error of only 1.54 is very good for a simple model that is not based on an exhaustive analysis.

CONCLUSIONS

- There are many stochastic processes observed in the field of water resources such as rainfall, water demand, river flows, evapotranspiration, and so on.
- A physical model, which can incorporate all of the uncertainties of such uncertain variables and subsequently accurately forecast is almost impossible to be developed.
- The theory of time series provides an excellent framework for the development of accurate models for stochastic processes. In time series context, auto-regrassive-integrated-moving average (ARIMA) model which are most powerful, accurate, and workable, give reasonably dependable forecasts.

- A monthly discharge sequance of River Indus observed at kalabagh is used as an example for the development of an ARIMA model. The ARIMA model is found to be a plausible model for such a time series.

- The forecast from the ARIMA model successfully track the seasonal pattern of the

data and the forecasts errors are very small.

- The ARIMA models can easily be developed by a suitable Computer package. Such models have a great potential for reflecting the effects of complex conditions and generating a dependable level of accuracy with a lesser amount of computing efforts.

- Forecasts from an ARIMA model can be very useful for the prediction of water demands and such forecasts can aid in the optimum management of any irrigation system in general and a demand irrigation system in particular.

REFERENCES

1. BOX, G.E.P. and Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, CA.

2. Cryer, J.D., 1986. Time Series Analysis, Duxbury Press, Boston, MA.

3. Hameed, T., Marino, M.A., and Shumway, R.H., 1994. Transfer function-noise modeling of evapotrnspiration. Jour. of Irrigation and Drainage Engineering, American Society of Civil Engineers(ASCE). Accepted for publication.

4. Pankratz, A., 1988. Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. Jhon Wiley and Sons, New York, NY.

5. Shumway, R.H., 1988. Applied Time Series Analysis. Prentice-Hall, Englewood Cliffs, NJ.