20th WEDC Conference Colombo, Sri Lanka, 1994

AFFORDABLE WATER SUPPLY AND SANITATION

Generation and forcasting of monsoon rainfall data

Mrs. D. Achela K. Fernando and A.W. Jayawardena, Hong Kong.

LONG TERM HISTORICAL records of hydrological information such as rainfall and runoff data form the basis of planning and design of major water resources projects. However, in most instances such historical records are often unavailable, and in situations where they are available, the records are too short to give any statistically significant meaning. One approach adopted to overcome this difficulty is to generate long term data synthetically. In this study, the outcome of an attempt to generate synthetic rainfall data for a station is presented. In the study, various stages of decomposing and synthesizing a time series are described and applied to a monthly rainfall data series from Sri Lanka. The results show that generated data preserve the basic statistical properties of the original series.

In the second part of the study, one step ahead forecasts are made using a Box-Jenkins (Box and Jenkins, 1976) type difference model. The results appear to be satisfactory.

Background

Time series analysis has been applied to many situations in the recent past. Some of the applications in water related areas include stream flow modeling (Avinash and Chanshyam, 1988), event rainfall data generation in semi-arid climates (Janos et al, 1988), detection of climatic changes (Geoff, 1989), water quality analysis (Jayawardena and Lai, 1989), rainfall for storm flow assessment (Henderson, 1989), and water quality forecasting (Jayawardena and Lai, I and II, 1991).

In this study an attempt has been made to apply time series analysis to monthly rainfall records of Peradeniya Botanical Garden, Sri Lanka(7° 20' N, 80° 38' E). The chosen station, situated in the tropical belt, experiences two monsoons: North-East from December to February and South-West from May to September. Monsoon precipitations form a large part of the rainfall at this station.

The analysis was done by decomposing the time series into their constituents. In the process of decomposition, properties characteristic to the station were revealed. Synthetic data generated proved to preserve basic statistical properties of the original series.

Forecasts were also made and compared with the observed values over a test period. Results were satisfactory.

Method of analysis

Prior to any analysis it is essential to verify that the series is homogeneous and stationary. Homogeneity implies that the data in the series belong to one population and therefore have a time invariant mean. Non-homogeneity arises due to changes in the method of data collection and the environment in which it is done. Stationarity, on the other hand, implies that the statistical parameters of the series computed from different samples do not change except due to sampling variations.

Homogeneity

Homogeneity implies a time invariant mean and, therefore, tests to check its prevalence are based on evaluating the significance of the change in mean value. A detailed account of the application of homogeneity tests to annual rainfall data in Hong Kong is presented by Jayawardena and Lau (1990) where Von Neumann Ratio, Cumulative Deviation and Bayesian Statistics have been used as test statistics.

Von Neumann Ratio (NR) is a statistic which has an expected value of 2 for homogeneous series. For non homogeneous cases it tends to get smaller than 2. The critical values are given by Owen (1962). Cumulative deviation measures the diviation from the mean by way of two statistics Q and R whose critical values are given by Buishand (1982). Bayesian statistics are measures of the change in mean value represented by statistics U and A, the critical values for which are also given by Buishand (1982). These statistics are also proportional to the departure from homogeneity.

Stationarity

A series, once found to be homogeneous, should also be tested for its stationarity. Because strict stationarity is only a mathematical concept, it is often necessary for practical purposes to restrict the conditions of stationarity to the mean and the variance only. Therefore the series is divided into several sub-series and the statistical characters of each sub-series are compared with those of the original series.

A homogeneous, stationary series can be represented as a linear combination of a trend component, a periodic component, a dependent stochastic component and an independent residual component.Time series analysis involves decomposing the series into its constituents.

Trend Component

Among many tests available to detect trend in time series are the Turning Point Test and Kendall's Rank Correlation Test (Kottegoda, 1980).

The Turning point test is based on the fact that too many or too few turning points indicate non-randomness. Kendall's Rank Correlation test statistically evaluates the trend of a series by computing the number of times p in all pairs of observations $x_i, x_j (j > i)$ that $x_j > x_i$. The theoretical concepts and an application of these tests is presented by Jayawardena and Lai (1989).

If a trend is detected it can be removed by fitting a polynomial function.

Harmonic analysis

If a periodicity exists in a trend free series, it can be removed by representing as a Fourier series of the form,

$$m_t = \mu + \sum_{i=1}^{h} [A_i \text{ Cos } (2\pi i\tau / p) + B_i \text{ Sin } (2\pi i\tau / p)]$$

where mt = the harmonically fitted means at period τ (τ = 1,2,3,...p); μ = the population mean; h = the total number of harmonics (= p / 2 or (p+1) / 2 depending on whether p is even or odd); p = period; A_i and B_i are the Fourier coefficients which are defined as,

$$A_i = (2/p) \sum_{\tau=1}^{p} \bar{x}_\tau \text{ Cos } (2\pi i\tau / p) \quad i = 1,2 \ldots .h$$

$$B_i = (2/p) \sum_{\tau=1}^{p} \bar{x}_\tau \text{ Sin } (2\pi i\tau / p) \quad i = 1,2 \ldots .h$$

where $\bar{x}t = (p/N) \sum_{\tau=1}^{N/p} x_{\tau + p(i-1)}$

For monthly data p = 12 and h = 6. Though it is not necessary to expand the series upto the maximum number of harmonics, in this study all 6 harmonics have been included.

The remaining part resulting from the separation of periodic component from a trend free series is the stochastic component.

Auto regressive and moving average models

Stochastic component is standardized by dividing by the periodic standard deviations. The dependent part of the stochastic component can be represented by an Auto Regressive Moving Average model, generally denoted by ARMA(p,q) where p and q are the orders of the autoregressive and moving average models respectively. General ARMA(p,q) model can be expressed as,

$$Z_t = \sum_{i=1}^{p} \phi_i \, Z_{t-p} + \eta_t + \sum_{i=1}^{q} \theta_i \, \eta_{t-q}$$

where h_t and Z_t denote the independent and dependent components respectively, f_i and q_i denote the auto regressive and moving average coefficients respectively.

The independent component can only be fitted by a probability distribution.

The optimum model fitting a stochastic series is that with the least number of parameters. This property of parsimony is tested by Akaike Information Criterion (Akaike, 1974) , the value of which takes a minimum for the optimum model.

Also, the residual should be independent. It can be tested by the auto-correlogram of the residual series and the Porte Mantau Lack of Fit test which are described in Jayawardena and Lai (1989).

Generation and forecasting

Synthetic data can be generated starting from a sample of the residual series using the appropriate probability distribution and following the procedure adopted in decomposing the series in reverse order.

The forecasts are made using the Box-Jenkins type difference model (Box and Jenkins, 1976) the equation for which is,

$$Z^*_{t+l} = \phi_1 [Z_{t+l-1}] + \ldots\ldots\ldots\ldots + \phi_p [Z_{t+l-p}] - \theta_1 [h_{t+l-1}] - \ldots$$
$$\ldots\ldots\ldots - \theta_q [\eta_{t+l-q} + [\eta_{t+l}]$$

where Z^*_{t+l} is the forecast at origin t for lead time l; Z_{t+l-i} is the observed sequence; h_{t+l} is the residual sequence. Square brackets [ ] indicate the conditional expectations which can be expressed as follows:

$[Z_{t-j}] = Z_{t-j}$, $j = 0,1,2,\ldots$; $[Z_{t+j}] = Z^*_t(j)$, $j = 1,2,\ldots$;

$[\eta_{t-j}] = Z_{t-j} - Z^*_{t-j-1}(1)$, $j = 0,1,2\ldots$; $[\eta_{t+j}] = 0$, j=1,2..;

One step ahead forecasts were made updating the data for the previous time level as they become available.

Data analysis and results

The monthly rainfall records used in this study are for the Peradeniya Botanical Gardens (1891-1981). The continuous record was split into two sub-series following preliminary and homogeneity tests the results of which are given in Tables 1 and 2. Table 2 shows that NR becomes smaller than 2, which indicates non-homogeneity, when the full series is considered and that it reaches values closer to 2 when sub-series are considered. Also, all statistics defined for Cumulative diviation and Bayesian statistcs assume larger values compared to their critical ones for the whole series indicating non-homogeneity and lie below them when considered as sub-series. This shows that the latter part of the series at the Botanical Garden form a different population. The low mean value for the latter part of the records has also been previously noticed (Fernando and Chandrapala, 1992).

After splitting into sub-series, each of them was further divided into two periods: a model development period and a model testing period (Table 3).

The sub-series did not show any significant trend. All six harmonics were fitted to the series. The harmonically fitted monthly means for the series are shown in Fig. 1. The twice a year periodic characteristic is revealed.

The resulting stochastic components were fitted with various ARMA models. The models tested were

Table 1. Basic Statistical Properties of Rainfall data Series.

Period	Mean	Std. Dev.
1891-1980 (All 92)	186.3	131.90
1891-1920 (First 30)	191.8	136.98
1921-1950 (Next 30)	192.7	129.76
1951-1981 (Last 32)	175.3	128.30

Table 2. Homogeneity test statistics for rainfall data at the Botanical Garden.

Period	NR	Q/√N	R /√N	U	A
1891-1980 (92)	1.8549 (n.a)	1.7765 (1.287)	1.6693 (1.659)	0.689 (0.456)	4.000 (2.48)
1921-1980 (60)	1.7041 (1.581)	1.8899 (1.274)	1.8899 (1.564)	1.081 (0.453)	5.585 (2.48)
1891-1960 (70)	2.1762 (n.a)	0.7952 (1.278)	0.2299 (1.578)	0.075 (0.454)	0.500 (2.48)
1961-1980 (22)	2.2067 (1.329)	1.0698 (1.224)	0.551 (1.444)	0.103 (0.447)	0.626 (2.44)

Note- Critical values for 95% confidence are in brackets.

Table 3. Sub-series of Rainfall data.

Series	Development	Comparison
BG-Series I	1891-1957 (67 Yrs)	1958-1960 (3 Yrs)
BG-Series II	1961-1978 (18 Yrs)	1979-1981 (3 Yrs)

Note:- BG - Botanical Gardens, SL.

AR(1),MA(1),ARMA(1,1),AR(2), MA(2),ARMA(2,1), ARMA(1,2),AR(3) and MA(3). Subroutines from the library package IMSL were used to determine the model parameters.

Table 4 shows the best models chosen according to the Akaike Information Criterion and the probability distribution of the remaining independent residual component.

Table 5 shows the basic statistical properties of the original and synthetic data. The hypothesis that the mean values of the generated data are not significantly different from those of the original data can be accepted at the 5% significance level.

One step ahead forecasts made using the Box-Jenkins type difference model are shown in Fig. 2 &3.

Table 4. Best models fitting the stochastic component.

Subseries	Model	Residual Distribution
BG- Series I	AR (3)	Log - Normal
BG-Series II	MA (1)	Log - Normal

Table 5. Statistical properties of Synthetic data.

Series	Original Data		Synthetic Data	
	Mean	Std. Dev.	Mean	Std. Dev.
BG- I	193.49 (184.2-202.8)	134.34	195.30	143.20
BG- II	166.64 (150.0-182.9)	123.51	167.88	121.57

Note:- BG =Botanical Gardens. Units are mm.
The 95% confidence limits are shown in brackets.

Conclusion

An attempt has been made to generate synthetic rainfall data for a station. Characteristics of the seasonality has been revealed during decomposition of the series. Models for data generation and forecasting have been developed. Generated data preserve the basic statistical properties of the original series. One step ahead forecasts are satisfactory except for a few peak occurrences in the testing periods.

References

Avinash, A. and Ghanshyam, D. (1988). Time Series Model of stream flow for a catchment of Ramganga River, J. Institution of Eng. India, Civil Eng. Div.,Vol.88 Part CI, pp 228-230.

Akaike, H. (1974). A new look at the statistical model identification, IEEE Transactions, Autom. Control, 1(6),716-723.

Box, G.E.P. and Jenkins, G.M. (1976). Time series analysis: forecasting and control. Holden Day, San Francisco, Calif.

Buishand, T. A. 1982. Some methods for testing the homogeneity of rainfall records. J. of Hydrology 58: pp11-27.

Fernando, T.K. and Chandrapala, L. (1992) Global Warming and rainfall variability- The Sri Lankan Situation, 5th Int. Meeting on Statistical Climatology, Toronto, Canada, pp 123-126.

Geoff, K. (1989) Use of Time Series Analysis to detect climatic change, J. of Hydro. Vol 111, pp 259-279.

Henderson, R.J. (1989). Rainfall Time Series for storm overflow assessment, Water Sci. Tech. Vol(21), pp 1789-1791.

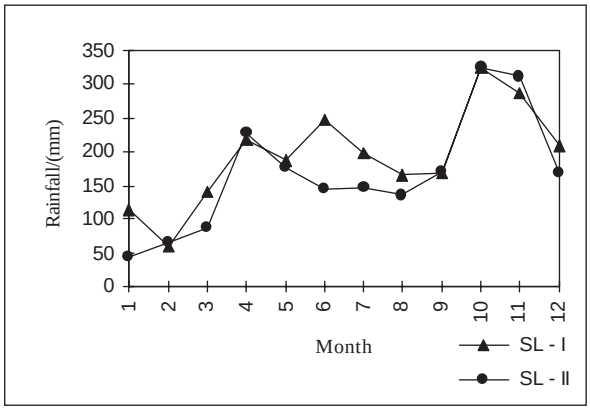

Figure 1. Harmonically fitted means for the rainfall series at Botanical Garden, Sri Lanka.

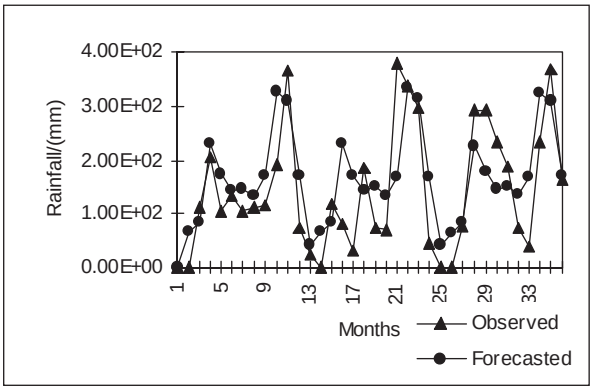

Figure 2. One step ahead forecasts for Botanical Garden 1979-1981

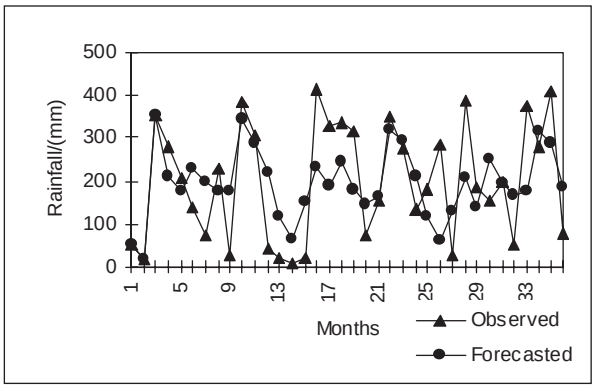

Figure 3. One step ahead forecasts for Botanical Garden 1958-1960

Janos, B., Lucien, D. and Omar, H. R. (1988) Practical generation of synthetic rainfall event time series in a semi-arid climatic zone, J. of Hydro., Vol 103, pp 357-373.

Jayawardena, A. W. and Lai, F. (1989) Time Series Analysis of water quality data in Pearl River, China, J. of Env. Eng., ASCE, Vol 155, No 3, pp590-607.

Jayawardena, A. W. and Lai, F. (1991 -I) Non-Gaussian ARMA modelling and forecasting of water quality data in Hong Kong, Proc. of the Int. Symp. On Env. Hydra., Hong Kong, pp1147-1153.

Jayawardena, A. W. and Lai, F. (1991 - II) Water quality forecasting using an adaptive ARMA modelling approach, Proc. of Int. Symp. on Env. Hydra., Hong Kong, pp 1121-1127.

Jayawardena, A.W. and Lau, W.H. (1990). Homogeneity tests for rainfall data, Hong Kong Engineer, Sept, pp 22-25.

Kottegda, N.T. (1980) Stochastic water resources technology, John Wiley, New York.

Owen, D.B. 1962. Handbook of statistical tables. Addison Wesley, Reading, Mass.