Jour. Ind. Soc. Ag. Statistics
47(3), 1995 : 278–287

Predicting the monthly rainfall around Guwahati using a Seasonal ARIMA model

D. K. Borah and P. K. Bora
Gauhati University
(Received : June, 1995)

SUMMARY

Monthly rainfall at Guwahati is modelled using a Seasonal ARIMA series. The model parameters are estimated using Marquardt algorithm for non-linear optimization. The various stages of model building are presented in a simple algorithmic form. The model is used to predict rainfall for the month ahead and monthwise rainfall for the year ahead.

Key Words : Seasonal ARIMA model, acf, pacf, periodogram, white noise, Marquardt Algorithm, Kolmogorov-Smirnov test.

Introduction

Rainfall plays the most important role in the agricultural economy of Assam. Almost all the total available supply of water for growing crops in the state is met by rainfall. Though heavy monsoon rainfall occurs every year in Assam, it varies from season to season, year to year and place to place. For example, the mean annual rainfall during the years 1951-1976 was 181.24 cm in Kamrup district with a standard deviation of 53.48 cm, whereas the same was 212.99 cm in Sibsagar district with a standard deviation of 29. 36 cm. There have been statistical studies [2], [3], [4] on the rainfall in India and its geographical regions. Parthasarathy and Mooley [2] have constructed a summer monsoon (June to September) rainfall series for India as a whole for the period 1866–1970. On the basis of application of Eisenhart's run test and Mann-Kendall rank statistic test, they have found that the series neither shows any significant trend nor any significant oscillations. Power spectrum analysis of the series has been reported to indicate a weak Quasi Biennial Oscillation (Period 2.3 to 2.8 years).

While most of the studies deal with the annual rainfall, we attempt modelling the monthly rainfall pattern of Guwahati using a Univariate Box - Jenkins (UBJ) model and predict the rainfall ahead of one month and one year. This study can be useful in predicting the occurrence of flood, planning of irrigation schemes and crop plantation in the agricultural belt around

REFERENCES

[1] Hader, R. J. and Park, S. H., 1978. Slope rotatable central composite designs, *Technometrics*, **20**, 413–417.

[2] Victorbabu, B. Re. and Narasimham, V. L., 1991a. Construction of second order slope rotatable designs through balanced incomplete block designs, *Commn. Statist.*, Theory and Methods, **20**, 2467–2478.

[3] Victorbabu, B. Re. and Narasimham, V. L., 1991b. Construction of second order slope rotatable designs through a pair of incomplete block designs, *J. Ind. Soc. Ag. Stat.*, **43(3)**, 291–295.

[4] Victorbabu, B. Re. and Narasimham, V. L., 1993. Construction of three level second order slope rotatable designs using balanced incomplete block designs, *Pak. J. Statist*, **9**, B, 91–95.

[5] Victorbabu, B. Re. and Narasimham, V. L., 1994. A new type of slope rotatable central composite designs, *J. Ind. Soc. Ag. Stat.*, **46(3)**, (to be appear).

Guwahati. In addition, it can be helpful in planning water supply schemes and drainage facilities for the city.

In the next three sections we discuss the theoretical and computational aspects of the UBJ model. Subsequently, we present the fitting of the model with monthly rainfall data of Guwahati over a nine year period and compare the values given by the model with the actual data. The predicted results are also compared with the available tenth year data.

2. *The Seasonal ARIMA model*

The Univariate Box-Jenkins (UBJ) model [1], often referred to as ARIMA (Auto-Regressive Integrated Moving Average) model, is one of the important and useful tools for time series modelling. This model is very convenient for both analysing the data sequence and forecasting.

Consider a non-stationary time series $\{Z_t\}$ whose d-th difference $\{\nabla^d Z_t\}$ is stationary. In this case, an ARIMA (p, q) model can be fitted to the differenced series as

$$\varphi_p (B) \nabla^d Z_t = \theta_q (B) e_t \tag{2.1}$$

where

$$\varphi_p (B) = \sum_{i=0}^{p} a_i B^i, a_0 = 1 \tag{2.2}$$

and

$$\varphi_q (B) = \sum_{i=0}^{q} b_i B^i, b_0 = 1 \tag{2.3}$$

with B being the backward shift operator and $\{e_t\}$ a white noise process.

The model (2. 1) is known as ARIMA (p, d, q) model and can represent a variety of physical situations.

When the time series contains seasonality with some period s, there are strong correlations between observations Z_t, Z_{t-s}, Z_{t-2s}, ... , and the series can be modelled by Box-Jenkin's multiplicative seasonal ARIMA $(p, d, q) \times (P, D, Q)^s$model as

$$\psi_p (B) \varphi_p(B^s) \nabla^d \nabla_s^D Z_t = \nu_q (B) \theta_Q (B^s) e_t \tag{2.4}$$

where $\psi_p(B)$ and $\nu_q (B)$ are the AR and MA operators as in (2.2) and (2.3) respectively and $\nabla_s^D = (1 - B^s)^D$ is the seasonal difference operator.

To speak in simple terms, the seasonal ARIMA model takes care of both trend and seasonality of the time series.

3. *Model Identification, parameter estimation and diagnostic checks*

Specifying the UBJ model involves model identification and parameter estimation. The model is accepted only after the diagnostic checks, verifying model adequacy.

3.1 Model identification

Model identification in the present case involves finding the appropriate values for s, d, D, p, q, P and Q. Identification requires the help from the non-parametric specifications in terms of statistical properties of the time series. These properties can be characterised by the autocorrelation function (acf) and partial autocorrelation function (pacf) in the time domain and the periodogram in the frequency domain [5].

3.2 Parameter estimation

The UBJ method favours estimates according to the Maximum Likelihood Estimation (MLE) criterion. But a Least-Squares Estimator (LSE) is easier to implement. It is to be noted that if the sequence $\{e_t\}$ is independent gaussian, the LSEs are asymptotically MLEs [5, 6].

Considering the equation (2.4), we get,

$$e_t = \theta_Q^{-1}(B^s)\, \nu_q^{-1}(B)\, \psi_p(B)\, \varphi_p(B^s)\, \nabla_s^D\, \nabla^d\, Z_t \qquad (3.2.1)$$

The least-squares estimators of the parameters in this case can be obtained by minimising $\sum_{t \in \{1, 2, \ldots, N\}} e_t^2$ with respect to the parameter set. We apply the Marquardt algorithm to find out the least-squares estimator. Marquardt algorithm is a powerful iterative technique that combines the good features of the Gauss-Newton method and the steepest descent method [7].

3.3 Model diagnostic checks

The diagnostic checks are performed to study the model adequacy. These checks include the study of the correlogram and the cumulative periodogram of the residual series. If the model fails the diagnostic tests, a new and improved model should be tried and the stages of identification, parameter estimation and diagnostic checks should be repeated.

3.3.1 Correlogram of the residual series

For the model adequacy, the whiteness property of the residuals must be tested. This means that in the correlogram, all the acfs except at lag 0 must lie within the ± 2SE (standard error) limits at 95% confidence level.

3.3.2 Cumulative periodogram

The cumulative periodogram is a very effective test for randomness of a series. It checks whether all the periodic characteristics of the series have been adequately taken into account. If some of the cumulative periodogram points lie above the upper significant line, presence of low frequency components are implied and if some points lie below the lower significant line, presence of higher frequency components are indicated [1,8].

The tasks of model identification, parameter estimation and diagnostic checks are presented in Algorithm 3.1.

4. *Forecasting*

Forecasting for a lead time m by UBJ method is based on finding the conditional expectation

$$\hat{Z}_{t+m} = E\left[Z_{t+m} / Z_t, Z_{t-1} \ldots \right] \tag{4.1}$$

Consider the multiplicative seasonal ARIMA model of equation (2.4). We can rewrite to get

$$Z_t = \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \ldots . + \beta_{p'} Z_{t-p'} + e_t + \\ + \mu_1 e_{t-1} + \mu_2 e_{t-2} + \ldots . + \mu_{p'} e_{t-p'} \tag{4.2}$$

where $p' = (P + D)\,s + p + d$ and $q' = Qs + q$. β_is and μ_is are functions of the seasonal and nonseasonal AR parameters and MA parameters respectively. Note that some of these parameters may be zero.

From (4.1) and (4.2) we get

$$\hat{Z}_{t+m} = \beta_1 \hat{Z}_{t+m-1} + \beta_2 \hat{Z}_{t+m-2} + \ldots . + \beta_{p'} \hat{Z}_{t+m-p'} + \\ + \mu_1 \hat{\theta}_{t+m-1} + \mu_2 \hat{\theta}_{t+m-2} + \ldots . + \mu_{q'} \hat{\theta}_{t+m-q'} \tag{4.3}$$

where $\hat{Z}_{t+m-j} = Z_{t+m-j}$ if $j \geq m, j = 1, 2, \ldots, p'$

and $\hat{\theta}_{t+m-j} = 0$ if $j < m$

$= \theta_{t+m-j}$ if $j \geq m, j = 1, 2, \ldots, q'$

Equation (4.3) provides the forecast function.

5. *Numerical results*

The monthly rainfall data for the period 1956 to 1965 are used here. The data are taken from the records of the hydrometric cell, Indian Meteorological Department, Guwahati, recorded at Guwahati Airport.

5.1 Model fitting

The data of the first nine years are used in the fitting of the model. The trace of the series (Fig. 5.1) indicates clearly a periodicity of 12 months. This is supported by the correlogram (Fig. 5.2) and the periodogram (Fig. 5.3) of the original series. The periodogram shows a prominent peak corresponding to a period of 12 time units.

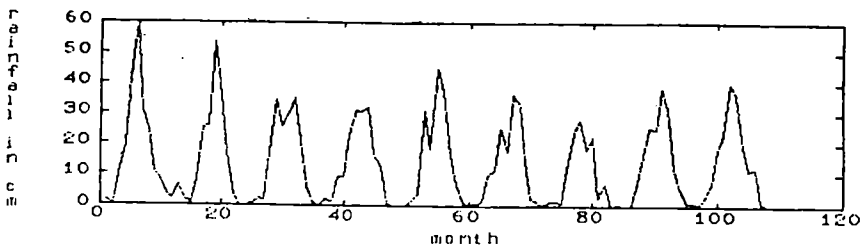

Fig. 5.1. Monthly rainfall (1956–1964)

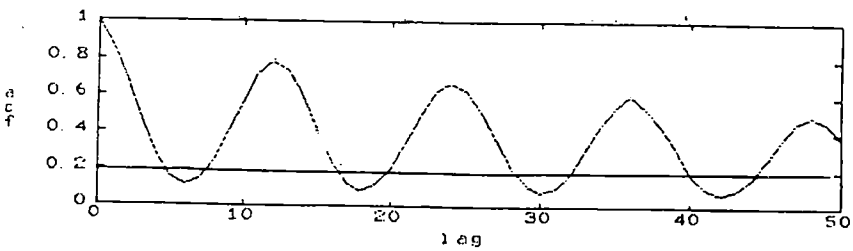

Fig. 5.2. Correlogram of rainfall series

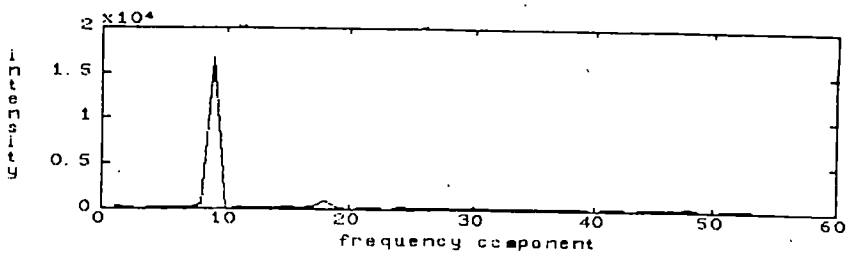

Fig. 5.3. Periodogram of rainfall series

We next observe the correlogram of the differenced series $\nabla_{12} Z_t$ (Fig. 5.4). Significant acf values are observed at the lags 3, 12 and 36. All other acfs in the figure are within $\pm 2SE$ and so it can be concluded that the seasonal differenced series with D=1 has become approximately stationary. The pacf diagram (Fig. 5.5) shows significant values at 9, 12 and 24. This suggests a tentative seasonal ARIMA (0, 0, 3) x (2, 1, 0)^{12} model as

$$Z_t = (1 + \varphi_1) Z_{t-12} - (\varphi_1 - \varphi_2) Z_{t-24} - \varphi_2 Z_{t-36} + e_t - \nu_1 e_{t-1} - \nu_2 e_{t-2} - \nu_3 e_{t-3} \quad (5.2.1)$$

The parameters are estimated as explained in section 3.2 and given in Table 5.1.

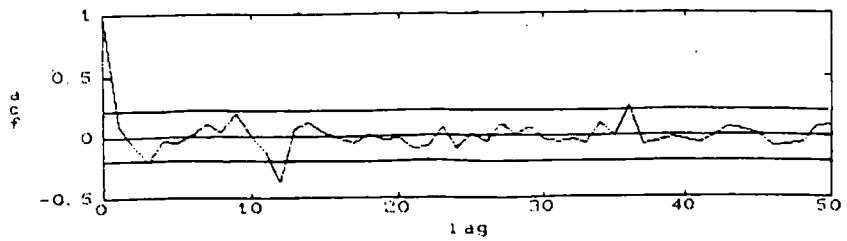

Fig. 5.4. Correlogram of ($\nabla_{12} Z_t$)

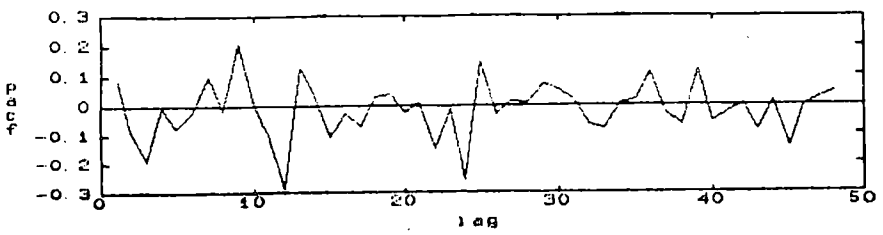

Fig. 5.5. Pacf plot of ($\nabla_{12} Z_t$)

Table 5.1. Estimated parameters of the seasonal ARIMA model

Parameters	Estimated values
Φ_1	–0.64670
Φ_2	–0.36059
ν_1	–0.18085
ν_2	–0.11607
ν_3	0.05060

The mean and the variance of the residual series are estimated as mean = – 0.35065 cm and $\sigma_\theta^2 = 27.47296$ cm^2.

The correlogram of the residuals (upto a lag of 30) in Fig. 5.6 shows that all the acfs are within $\pm 2SE$ limits. Further from the cumulative periodogram of the residual series (Fig. 5.7) it is observed that all the points lie within the two Kolmogorov-Smironov 95% confidence limit boundaries. Hence the residual series is white. The patterns in the data are, therefore, duly taken care of and the model can be used for forecasting.

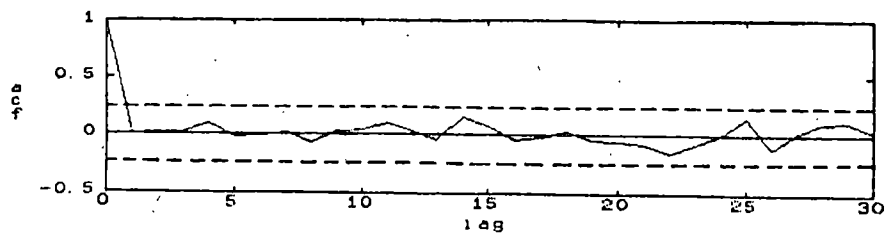

Fig. 5.6. Correlogram of residual series

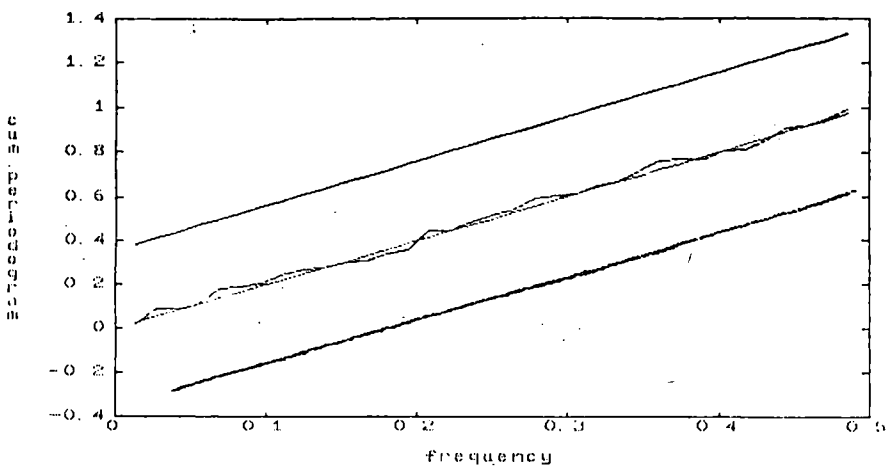

Fig. 5.7. Cumulative Periodogram

5.2 Forecasting

Using the model (5.2.1), one month ahead forecasts and forecasts for lead time m=1, 2, ... , 12 are made at the time origin t = 108. Table 5.2 gives a comparison of the forecasts with the observed values. A part of the fitted model and forecasts are shown in Fig. 5.8 and Fig. 5.9 along with the observed values.

Table 5.2. Comparison of the observed data and forecasts

Time t (month)	Observed values (cm)	Forecasts (cm)	
		One step	12 time units
109	0.00	0.00	0.00
110	3.90	1.25	1.20
111	7.50	5.19	4.67
112	19.60	16.28	15.56
113	31.80	25.05	24.32
114	36.30	32.84	31.35
115	37.00	31.56	30.32
116	28.30	25.00	23.96
117	15.00	9.34	8.29
118	2.60	9.33	8.20
119	2.80	0.00	0.52
120	0.20	0.00	0.30

Discussion

1. From Figs 5.8, 5.9 and table 5.2 it is clear that the model well represents the rainfall pattern over a range of rainfall values which vary from 0 cm to more than 50 cm and give reasonably good predictions.
2. Since the rainfall in some months in the winter is scanty (sometimes even 0 cm is a month) the predictions for such months are not good.
3. As discussed previously there may be year to year periodic variation of rainfall data. This variation should also be taken into account for modelling and prediction. However, such a study requires rainfall data for a sufficient number of years and is beyond the scope of the present work.

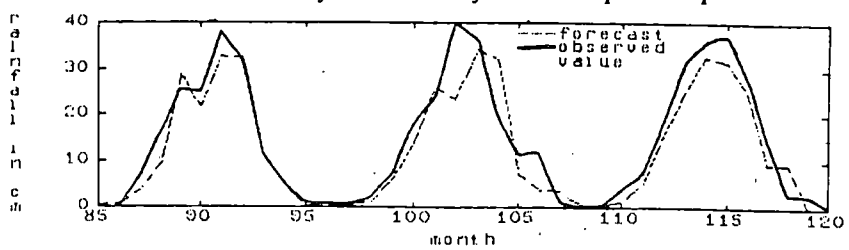

Fig. 5.8. Model and one-step-ahead forecast

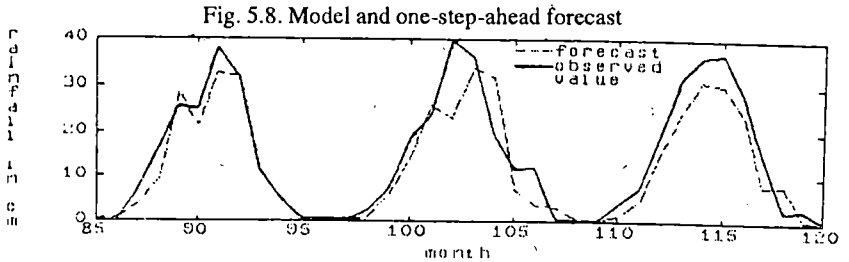

Fig. 5.9. Model and twelve-step-ahead forecast

Conclusion

We have fitted a seasonal ARIMA model to the monthly rainfall data of Guwahati and estimated the model parameters. The model is found to adequately represent the series. The forecasts made by the model are compared with the observed values and are found to be reasonably good. The algorithm presented here is general in nature and can be adopted to model monthly rainfall data of any place.

Algorithm 3.1

Given the rainfall series X(t) for t = 1, ..., N.

Initialize the orders of difference d = 0 and seasonal difference D = 0.

Step 1 Estimate the mean of the series and subtract it from X(t) to get the modified series Z(t).

Step 2 Consider d-th difference of the series.

Plot trace, acf, spectrum of the differenced series.

Step 3 Check for seasonality.

Step 4 If seasonality is present consider the next value of D.

Plot trace, acf, pacf, spectrum of the series and go to step 3.

Step 5 Test for stationarity.

If approximate stationarity is achieved,

then select those values of d and D

else select new value of d and go to step 2

Step 6 Compare with theoretical acf, pacf, patterns of common ARIMA processes. Suggest structures.

Step 7 Select a new (d,D, P, Q, p, q) from the suggested structures.

Step 8 Make initial estimates of $\psi, \varphi, \nu, \theta$.

Step 9 Improve estimates.

Step 10 Compute model residues. Perform model diagnostic checks.
If the checks fail then go to step 7.

Step 11 The model is ready.

REFERENCES

[1] Box G. E. P. and Jenkins G. M., 1976. Time Series Analysis Forecasting and Control, Holden-Day Inc.

[2] Parthasarathy and Mooley, 1978. 'Some features of a long homogeneous series of Indian rainfall', *Monsoon Weather Review*, vol. **104**, no . 6, June issue.

[3] Mooley, D. A., 1992. 'The Indian Summer Monsoon, its Economic Aspects, Vagaries and Remedial Measures' in The Ecology of Agricultural Systems, (New Dimensions in Agricultural Geography, vol. 2), Noor Mohammad Ed., Concept Publishing Company, New Delhi.

[4] Das, M. M, 1992. 'Climatic pattern and Agro-climatic regions of Assam' in The Ecology of Agricultural System, Concept's International Series in Geography No. 4, Noor Mohammad Ed., Concept Publishing Company, New Delhi.

[5] Priestley M. B., 1981. Spectral Analysis and Time Series, Vol. I & II, Academic Press, N. Y.

[6] Mohanty N., 1986. Random Signals Estimation and Identification Analysis and Applications, Von Nostrand Reinhold Company.

[7] Marquardt D. W., 1963. 'An Algorithm for least squares estimation of non-linear parameters', *Journal of SIAM* **11**.

[8] Jenkins G. M. and Watts D. G., 1968. Spectral Analysis and its applications, Holden Day, San Francisco.

Jour. Ind. Soc. Ag. Statistics
47(3), 1995 : 288–304

Regional Disparities in the Levels of Development in Uttar Pradesh*

Prem Narain, S.C. Rai and Shanti Sarup
Indian Society of Agricultural Statistics, New Delhi - 110 012

SUMMARY

The level of development of various districts of Uttar Pradesh was estimated with the help of composite index based on optimum combination of thirty eight economic indicators. All the sixty three districts of the State were included in the study. The data for the year 1991-92 on thirty eight economic indicators were used. Eighteen indicators were directly concerned with agricultural development, seven indicators depicted the progress of development in industrial sector and the rest thirteen indicators presented the level of development in infrastructural service sector.

The level of development was examined separately for agricultural, industrial and overall socio-economic sectors. The district of Ghaziabad was found to rank first and that of Chamoli was the last in the overall socio-economic development. Wide disparities in the level of development has been observed among different regions of the State and the western region had been found to be better developed as compared to other regions of the State. Positive significant association was found between the levels of development in the agricultural and industrial sectors indicating that the growth and progress of agriculture and industry had been going hand in hand in the State. Six districts covering about 9 per cent area and little more than 10 per cent population of the State, were found to be better developed whereas twenty three districts having 41 per cent area and 35 per cent population were categorised as low developed districts.

For bringing about uniform regional development, potential targets for various indicators had been estimated for poorly developed districts. The study revealed that the low developed districts required improvements of various dimensions in most of the indicators for enhancing the level of overall socio-economic development.

Key words : Composite index, Development indicators, Model districts, Potential target, Regional disparities.

Introduction

Uttar Pradesh is primarily an agricultural state. The total foodgrains production of the State during 1989-90 was of the order of 338 lakh tonnes

* Study undertaken in the Research Unit of ISAS during 1995.