Nigerian Journal of Social and Development Issues, Vol. 3, No.1, January 2003
ISSN 1596-2288 © Copyright Faculty of Social Sciences, University of Calabar. Printed in Nigeria

FORECASTING RAINFALL IN CALABAR USING SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (SARIMA)

By

EZE BASSEY EZE
DEPARTMENT OF GEOGRAPHY AND REGIONAL PLANNING
UNIVERSITY OF CALABAR
CALABAR

And

CHRISTIAN E. ONWUKWE
DEPARTMENT OF MATHS/STATS. & COMPUTER SCIENCE.
UNIVERSITY OF CALABAR,
CALABAR

ABSTRACT

The available statistical analysis of rainfall data in Calabar is limited to minimum, maximum, mean monthly and annual totals. Further analyses have often appeared in the form of range, line and bar graphs. These simplistic analyses are not very useful in the prediction of future patterns of rainfall or in the evaluation of climate change and variability. The possibility of predicting future rainfall in Calabar is investigated using seasonal autoregressive integrated moving average (SARIMA) model which is fitted on a 10-year old rainfall data of Calabar (January 1991 – December 2000). The fitted model turned out to be $W_t = (1-0.95393B)(1-08790B^{12})a_t$. *The model was used as a basis for forecasting rainfall for Calabar town for a period of 4 years (2001 - 2004) and the results were found to be satisfactory. The model is therefore deemed suitable for not only forecasting rainfall for the town of Calabar but for other towns in the wet tropic.*

INTRODUCTION

Prediction is the most important aspect of applied climatology and only models can be used for this purpose. Prediction also known as forecasting is one area of the subject that is most fascinating to members of the public who utilize weather information. Apart from this, it is also clear that we will be making very little impact on the current debate on climate change without models. Yet models of weather or climatic patterns are lacking for most weather elements in the third world. The reasons for this are not far fetched. The first is the inadequacy of information available at the disposal of researchers (Ayoade, 1988). The second reason has to do with the dearth of climatologists and meteorologists (Adefolalu, 1982). The third is the seeming lack of interest in the subject of climatology by many persons including geographers. Yet the science of meteorology is as crucial to man as the air around us!

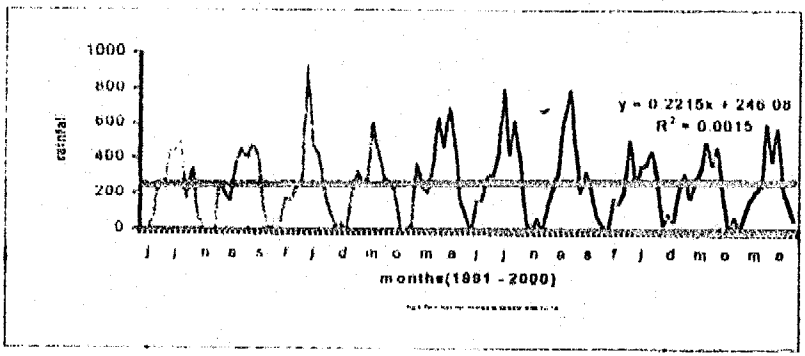
With all these reasons, it is therefore not surprising that the available weather statistics are limited and far in between. For example, the rainfall statistics available for a town like Calabar, the wettest town in Eastern Nigeria, are simply minimum, maximum, mean monthly and annual rainfall figures. Further analysis appear merely as bar and line graphs. This type of simplistic analyses are rather too broad for any serious scientific consideration. Elsewhere as in Europe, USSR and United States, rigorous statistical techniques including multivariate analyses have been employed. However, Akintola (1986) analysed the rainfall data for Ibadan using time series. In fact, time series has always been useful in modeling rainfall patterns. Usually, it involves the calculation of moving averages in which values are determined for successive overlapping periods of five, ten or thirty years (Barry, 1969). This type of analysis is usually the basis for prediction of rainy days, inception periods, cessation periods, confidence limits etc. In fact fitting a time series model usually reveals clear-cut patterns of cycle, randomness or bundiness (Akintola, 1986). Often the aim is to interpret past patterns of rainfall in terms of future probabilities.

METHOD

Table 1 shows the monthly rainfall totals in millimeters collected from Meteorological Centre, Calabar International Airport from January 1991 to December 2000. The problem is to fit a suitable model to the data and produce forecasts for up to four years ahead.

TABLE 1: MONTHLY RAINFALL TOTALLS FOR CALABAR (1991 – 2000)

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
J	1	2.6	64.3	40.7	1.2	2.4	61	25.4	86	66
F	11	0.5	39.9	0	21.1	162.2	0	6	49.8	0
M	88.3	271.5	177.7	167.3	366.1	161.5	139.4	174	223	95.9
A	259.3	206.3	161.2	328.2	248.3	314.5	228.7	149.1	311.4	166.4
M	258.4	161.8	230.1	255.3	208.6	299.5	328.2	211.6	180	217
J	479	387.7	281	294.2	375.2	435.3	633.1	504.5	270.3	250.6
J	439.6	455.1	927.4	609.5	634.2	801.1	796.6	255.2	349.9	597.9
A	505.6	400.2	479.1	424.5	467	425.4	492.6	353.7	494.5	392
S	169.8	481.4	420.5	290	696.8	615.2	211.2	365	368.3	577.6
O	359	408.5	244.6	265.2	496.1	377.4	319.1	439.1	463.7	232.9
N	72.2	120.7	119.7	229.7	168	40.5	214.3	296.3	207.2	153.6
D	19.7	0.1	22.8	0	69.7	0.6	68.2	33.6	0.3	57.5
Tota	2662.9	2896.4	3168.3	2903.9	3752.3	3635.6	3492.4	2813.5	3004.4	2697.4



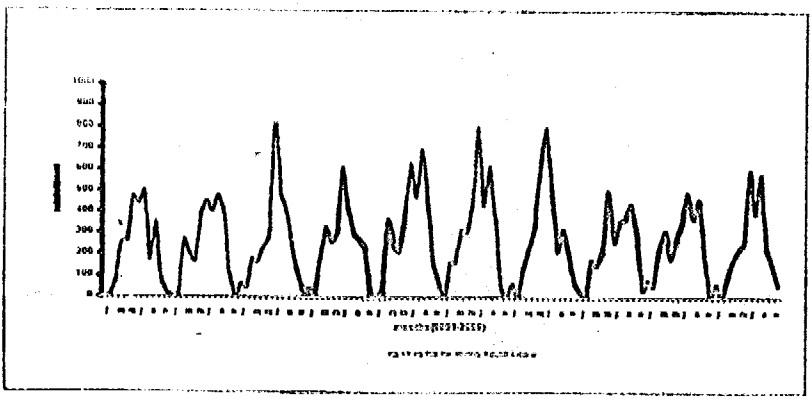


Figure 1 shows a clear seasonal pattern as well as slight upward trend.
The magnitude of the seasonal variation increases at the same sort of rate as the yearly mean levels and this indicates that a multiplicative seasonal model is appropriate. Following Box and Feukins (1970) and other authors, we will try to fit Autoregressive Integrated Moving Average (ARIMA) model. Since Figure 2 shows trend and the size of the seasonal effects appears to increase with the mean, then we may transform the data to make the seasonal effect additive by use of logarithmic transformation.
The simplest trend is the familiar "linear trend + noise", for which the observation at time t is a random variable x_t given by

$$x_t = \alpha + \beta_t + \varepsilon_t \quad \text{---------- (1)}$$

Where α, β are constants ε_t denoted a random error with zero mean The mean level at time t is given by $m_t = (\alpha + \beta_t)$. This sometimes called "trend term". The analysis of a series which exhibits trend depends on whether one wants to measure trend or remove the trend in order to analyse local fluctuations. It also depends on whether the data exhibit seasonality. Since our data exhibit seasonality, it is a good idea to start by calculating successively yearly averages as these will provide a simple description of the underlying trend.

A second procedure for dealing with a trend is to use a linear fitter which converts one time series $\{x_t\}$ into another $\{Y_t\}$ by the linear operation

$$y_t = \sum_{r=-q}^{+S} a_r x_{t+r} \quad \text{---------- (2)}$$

Where $\{a_r\}$ is a set of weights. In order to smooth out local fluctuations and estimate the local mean, we choose the weights so that $\Sigma a_i = 1$ and then the operation is often referred to as a moving average. Kendell, Stuart and Ord (1983). Moving average are often symmetric with $s = q$ and $a_j = a_{-j}$. The simplest example of symmetric smoothing filten is the simple moving average, for which $a_r = {}^1/_{(2q+1)}$ for $r = -q, \dots + q$ and the smoothed value of x_t is given by

$$S_m x_t = \frac{1}{2q+1} \sum_{r=-q}^{+q} x_{t+r} \quad \text{---------- (3)}$$

The simple moving average is not generally recommended by itself for measuring trend, although it can be useful for removing seasonal variation.

A special type of filtering, which is particularly useful for removing a trend, is simply to difference a given time series until it becomes stationary (no symmetric change in mean, no symmetric change in variance, and if strictly periodic variations

have been removed). This method is an integral part of the procedures advocated by Box and Jenkins (1970). For non-seasoned data, first-order differencing is usually sufficient to attain apparent stationarity, so that the new series $\{y_1, \text{-----}, y_{N-1}\}$ is formed from the original series $\{x_1, \text{-----}, x_N\}$ by

$$y_t = x_{t+1} - x_t = \nabla x_{t+1} \quad \text{----------} \quad (4)$$

First-order differencing is widely used. Occasionally, second-order differencing is required using the operator ∇^2

Three seasoned models in common use are

(I) $x_t = m_t + s_t + \varepsilon_t$

(II) $x_t = m_t\, s_t + \varepsilon_t$

(III) $x_t = m_t\, s_t\, \varepsilon_t$

Where m_t is the deseasonalized mean level at time t, s_t is the seasonal effect at time t, and ε_t is the random error. Model I describes the additive case while models II and III both involve multiplicative seasonality. In model III the error term is also multiplicative and a logarithmic transformation will turn this into a (linear) additive model which may be easier to handle. The time plot should be examined to see which model is likely to give the better description. The seasonal indices $\{s_t\}$ are usually assumed to change slowly through time so that $s_t \simeq s_{t-s}$, where s is the number of observations per year. The indices are usually normalized so that they sum to zero in the additive case, or average to one in the multiplicative case.

A seasoned effect can be eliminated by differencing (Box and Jenkin (1970)). For example with monthly date one can employ the operator ∇_{12} where

$$\nabla_{12} x_t = x_t - x_{t-12} \quad \text{----------} \quad \text{----------} \quad (5)$$

Alternative methods of seasoned adjustment are reviewed by Pierce (1980), Cleveland (1983) and Newbold (1984). These include the widely used x - 11 method which employs a series of linear filters.

Many series contain a seasonal periodic component. The seasonal effect implies that an observation for a particular month say, October is related to the observations for previous Octobers. With monthly observations $\{x_t\}$ where s = 12 may certainly expect x_t to depend on terms such as x_{t-12} and may be x_{t-24}, as well as terms such as x_{t-1}, x_{t-2}, ----. Box and Jenkins (1970) have generalized the Autoregressive Integrated Moving Average (ARIMA) to deal with seasonality, and define a generalized multiplicative Seasonal Autoregressive Integrated Moving Average (SARIMA) Model as

$$\phi_p(B)\,\Phi_P(B^s) W_t = \theta_q(B)\,\Theta_Q(B^s)\, Z_t \quad \text{----------} (6)$$

Where B denotes the backward shift operator, ϕ_P, Φ_P, θ_q Θ_Q are polynomial of order p, P, q, Q respectively, Z_t denotes a purely random process and

$$W_t = \nabla^d \nabla_s^D X_t \quad \text{----------} \quad \text{----------} \quad (7)$$

The variable $\{W_t\}$ are formed from the original series $\{x_t\}$ not only by simple differencing to remove trend, but also by seasonal differencing V_s to remove seasonality.

The model in equations (I) and (II) is said to be a SARIMA model of order (p,d,q) X (P,D, Q). The values of d and D do not usually need to exceed one. When fitting a seasoned model to data, the first task is to asssess values of d and D which reduce the series to stationarity and remove most of the seasonality. Then the values of p,P, q and Q need to be assessed by looking at the Autocorrelation functions (acf) and partial autocorrelation functions (pacf) of the difference d series and choosing a SARIMA model whose act and pact are of similar form. Finally, the model parameters may be estimated by some suitable iterative procedure.

RESIDUAL ANALYSIS

When model has been fitted to a time series, it is advisable to check that the model really does provide an adequate description of the data. This is usually done by looking at the residuals which are defined by

Residual = Observation – Fitted Value

If we have a "good" model then we expect the residuals to be "random" and "close to zero", and model validation usually consists of plotting residuals in various ways. With time series models we have the added feature that the residuals are ordered in time and it is natural to treat them as a time series.

The two obvious steps are to plot the residuals as a time, plot, and to calculate the acf of the residuals. The time plot will reveal any outlines (extreme values) and any obvious autocorrelation or cyclic effects. The residual correlogram will enable auto correlation effect to be examined more closely.

If $\{r_K\}$ denote the autocorrelation coefficient at log K of the $\{Z\}$. if we have fitted the true model then the true errors form a purely random process and their correlogram is such that each autocorrelation coefficient is r $r_K \sim N(0, {}^1/_N)$for reasonably large values of N.

Box and Jenkins (1970) describe what they call a portmanteau lack-of-fit test which looks at the first M values of the correlogram all at once. The test statistic is

$$Q = N \sum_{k=1}^{m} r_k^2 \qquad \text{(8)}$$

Where N is the number of terms in the differenced series and m is typically chosen in the rauge 15 to 30. if the fitted model is appropriate then Q should be approximately distributed as χ^2 approximation can be rather poor for N < 100 and various alternative statistics have been proposed. Ljung and Box (1978) suggest $\frac{N(N+2)\sum r_k^2}{N-K}$. However, the test have rather poor power properties Dallies and

Newbold (1979). A variety of other procedures have also been proposed for looking at residual, Newbold (1988) but our own preference is usually just to look at the few values of r_k particularly at logs 1, 2 and the first seasonal lag (if any and see if any are significantly different from zero using the limits of $\pm 2/\sqrt{N}$. If they, then modify the model in an appropriate way by putting in extra term to account for the significant coefficient(s). However, if only one (or two) values of r_k are just significant at lags which haye one obvious physical meaning, then there would not be enough evidence to reject the model.

DISCUSSION

First, we transform the data by taking the logarithm of the original data. Transformation does three things to a raw data – it helps to stabilize the variance, makes seasonal effect additive and finally makes the data normally distributed. Since model building and forecasting are usually carried out on the assumption that the data are normally distributed. The first 38 coefficients of the acf and pacf that is, Fig 3 and Fig. 4 of the logged data show pattern which suggest the presence of seasonal variations. Some sort of differencing is clearly required

With monthly seasonal data, the obvious operators to try are

$$\nabla, \nabla_{12}, \nabla\nabla_{12} \text{ and } \nabla^2_{12}$$

If $y_t = \log x_t$ denote the logarithms of the observed data, we start by looking at the acf and pacf of ∇y_t. With N observations in the differenced series, a useful rule-of-thumb for deciding if an autocorrelation coefficient is significantly different from zero is to see if its modulus exceeds $2/\sqrt{N}$. Hence the critical value is 0.183 and we find significant coefficients at lags 1, 2, 9, 12,24, 34, 35, 36. There is no sign that the acf if damping out, and further differencing is required. For $\nabla_{12}y_t$, the series is still nonstationary and so we next try $\nabla\nabla_{12}Y_t$. The number of terms in the differenced series is now 107, and an approximate critical value is 0 194. We note significant values at lags 1 and 3 but most of the other values are small and there is no evidence on non-stationarity. Thus we choose to fit an ARIMA model to $\nabla\nabla_{12}Y_t$. In

order to identify a suitable ARMA model for $\nabla\nabla_{12}Y_t$, we need to look at pacf in Fig. 6. As for the acf Fig.5, coefficients whose moduli exceed 0 194 may as a first approximation, be taken to be significantly different from zero. In pacf we note significant values at lags 1 and 3. When significant values occur at unusual lags they are usually ignored unless there is external information as to why such a lag should be important

We should now be in a position to identify an appropriate seasonal ARIMA model to fit to the data. This means that we want to assess values of P,q, P and Q in the model defined by equation (6). The seasonal values P and Q are assessed by

looking at the values of the acf and pacf at lags 12, 24, 36 ——. In this case the values are large at lag 12 but small at lags 24 and 36, indicating no autoregressive term but one seasonal moving average term. Thus we take P = 0 and Q = 1. The values of the non-seasonal values p and q are assessed by looking at the first few values of acf and pacf. The only significant values are at lags 1 and 3, and these values are not easy to interpret. An AR(3) model, as suggested by the pacf would generally have a slowly decaying acf, while the MA(3) model, suggested by the acf, would generally have a slowly decaying pacf. Noting that the coefficient at lag 3 are only just significant, we could as a first try, take just one moving average term and set p = 0 and q = 1. if we now work out the standard error of the autocorrelation coefficient at lag 3 using a more exact formula, we find that it is not in fact significantly different from zero. This more exact formular Box and Jenkins (1970) assumes that an MA(1) model is

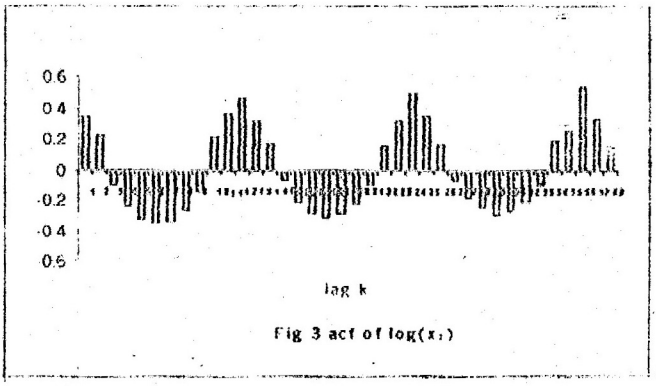


Fig 3 acf of $\log(x_t)$

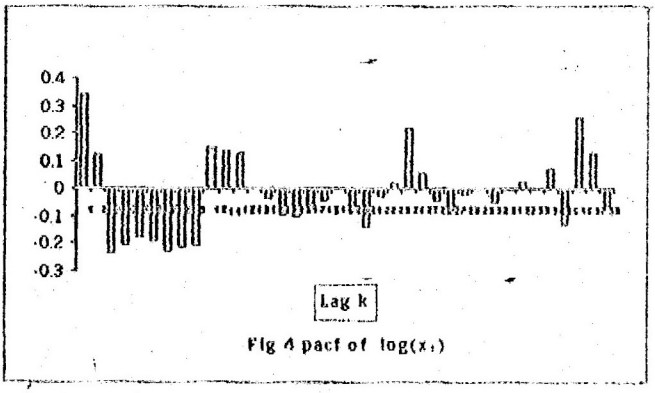


Fig 4 pacf of $\log(x_t)$

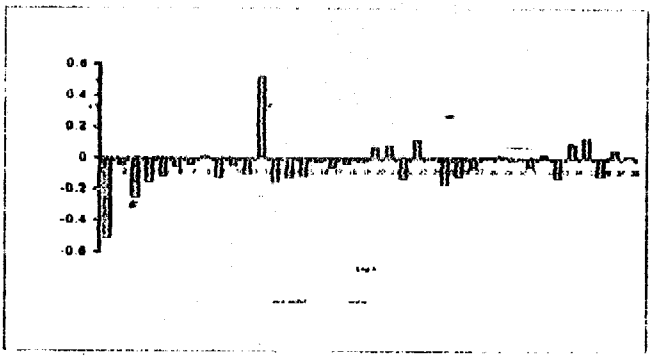
appropriate rather than a completely random series for which the formula $1/\sqrt{N}$ is appropriate. This result gives us more reliance on the choice of p = 0 and q = 1. Thus we fit a SARIMA model with p = 0, d =1, q = 1 and P = 0, D = 1, Q = 1 setting $W_t = \nabla\nabla_{12} \log X_t$, the fitted model terms out to be

$$W_t = (1-0.9539B)(1-0.8790B^{12})a_t$$

Fig. 7 and Fig. 8 show the time plot and acf of residuals respectively. Inspection of acf shows that none of the coefficient is significantly different from zero (although about 1 in 20 coefficient, will be significant at the 5 per cent level under the null hypothesis that the residuals are random). Also, a stem and leaf display shows somewhat normal distribution. There is no evidence that our fitted model is inadequate and so no alternative model will be tried.

FORECASTS

It was also necessary to attempt a forecasts of the rainfall which might occur in Calabar within a 4-year period. The result is presented below on table. From this result, it is clear that this model can be very useful in predicting rainfall in Calabar.



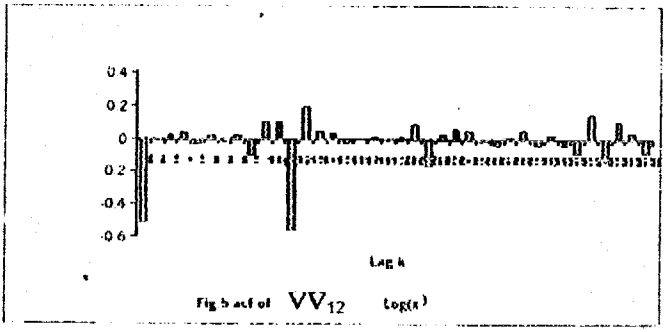


Fig 8 acf of $\nabla\nabla_{12}$ log(x)

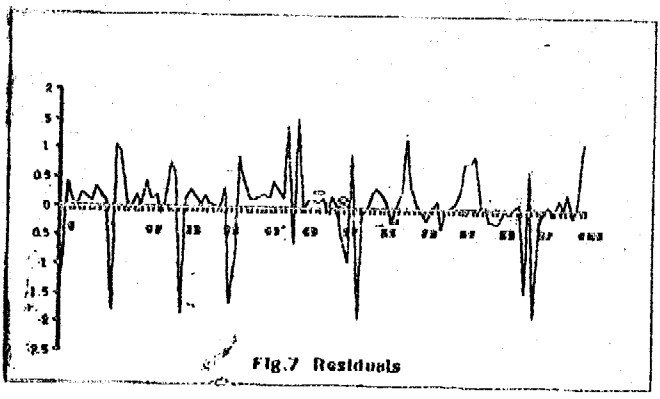


Fig.7 Residuals

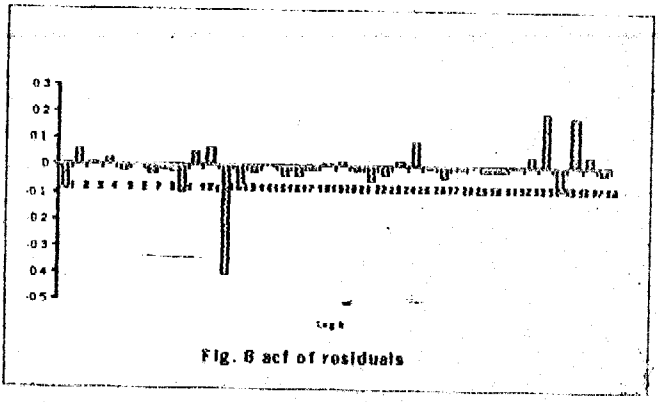


Fig. 8 acf of residuals

Table 2: Rainfall Predictions Using the SARIMA Model

Forecast

	J	F	M	A	M	J	J	A	S	O	N	D
2001	17.884	3.929	165.300	222.363	214.255	344.053	492.006	422.751	371.163	342.285	145.741	6.177
2002	17.454	3.834	161.324	217.014	209.101	335.777	480.171	412.582	62.235	334.052	142.235	6.028
2003	17.034	3.742	157.443	211.794	204.071	327.700	468.622	402.658	353.523	326.017	138.814	5.883
2004	16.624	3.652	153.657	206.700	199.163	319.819	457.351	392.975	345.020	318.176	135.476	5.742

From Table 2, it is clear that the model approximate reality. For example,

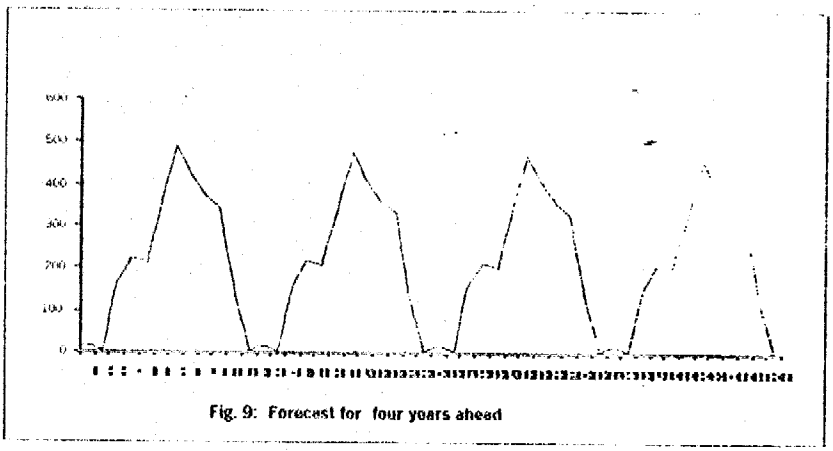


Fig. 9: Forecast for four years ahead

the forecast for the month of January shows that rainfall for the next four years may average 17 m and for December it may not exceed 6 mm.

CONCLUSION

Geographers and Environmentalists are interested in long-term variability of climatic parameters. This study has analysed a 10-year Rainfall data of Calabar using seasonal Autorogressive Integrated Moving Average (SARIMA). This fitted model turned out to be

$$W_t = (1-0.95393)\,(1-0.8290B^{12})a_t$$

This model was used as the basis for forecasting rainfall figures for Calabar for the next 4 years. It is our conclusion that this model will be very useful in the prediction of rainfall for not only Calabar but for other towns in the humid tropics. And information on predicted values of rainfall are very essential at this time of perceived global warming and climate change. Indeed we can only prepare for periods of marked deviations from the mean if we know and can predict such deviations (Hayward & Oguntoyuibo, 1987)

REFERENCES

Adefolalu, D. O. (1982) Personal Communication.

Akintola, J. O. (1986). *Rainfall Distribution in Nigeria 1892 1983.*

Ayoade, J. O. (1988) *Introduction to Climatology for the Tropics*. Spectrum Books, Ibadan.

Bary, R. C. (1969). *Introduction to Physical Hydrology*. Methuen.

Bartlett, M. Sc. (1966). *Stochastic Processes (2nd edn.)* Cambridge: Cambridge University Press.

Box, G.E.P. and Jenkins, G.M. (1970) *Time Series Analysis, Forecasting and Control.* Francisco: Holden-Day (revised Ed. Published 1976).

Chertfield, C. (1989). *The Anatysis of Time Series: An Introduction* (4th edn.) London, Chapman and Hall, (reprinted 1991. 1992).

Cleveland, W. S. (1983). *Sesonal and Calendar Adjustment: In Handbook of Statistics.* Vols. 3, (eds. D. R. brillinger and P. R. Krishniali), Amsterdam: North-Holland 39 -- 72.

Cleveland, W. S. and Derlin, S. J. (1982) *Calendar effects in Monthly Time Series: Measurement and Adjustment*, J. Amer. Stgaistics Ass., 77, 520- 8.

Dallies, N. and Newsbold, P. (979). Some power studies of a portmanteau text of time series model specification. Biometrika, 66, 153-5.

Kendall, M. G., Stuart, A. and Ord. J. K. (1983). *The Advanced Theory of Statistics.* Vol.3 (4th edn.) London Griffin.

Ljung, G.M. and Box, G. E.P. (1978). On a measure of lack of fit in in time series Models Biometrika, 65, 297 – 303.

Newbold, P. (1984) *Some recent development in two-series analysis, I, II and III Int. Statist. Rev., 49, 53 -- 66; 52, 183 -- 92, 56, 17 – 20.*

Pierce, D. A. (1980) *Some recent developments in Sesonal Adjustment* in Directions in Time-Series (eds. D. R. Brillinger and G. C. Tiao). Institute of Mathematical Statistics.

END NOTE

Adefolalu, D. O. (1982) Personal Communication, Calabar