Auto-Regressive Integrated Moving-Averages Model For Daily Rainfall Forecasting

Rashmi Bhardwaj, Varsha Duhoon

Abstract: Weather on earth changes continuously and hence becomes important to forecast so as to foresee any natural calamity and hence to take preventive measures in advance. Auto Regressive Integrated Moving Averages model (ARIMA model) is the time series model which studies the stationary series and hence is used for modelling and forecasting. This paper studies the ARIMA model in order to model and forecast the weather parameter as rainfall of Delhi region from January 1, 2017 till December 31, 2018 on daily basis. In order to check the stationarity of the time series, Augmented-Dickey-Fuller(ADF) and PhilipsPeron(PP) tests have been applied. The study revealed that the ARIMA(1,1,1) is suited model for the series on the basis of the AkaikeInformationCriterion(AIC) value used for choosing the appropriate model for forecasting parameter.

Key Words: Auto Regressive Integrated Moving Averages model (ARIMA model), Augmented Dickey-Fuller (ADF), Philips Peron (PP), Akaike Information Criterion (AIC)

1. INTRODUCTION

In Indian economy agriculture plays a significant role affecting the Gross Domestic Product (GDP) of the nation: agriculture alone makes twenty one percent contribution to GDP and provides sixty percent employment. Nearly forty percent of population in the country is supported by rainfed agriculture which occupies sixty seven percent of net sown area yielding forty four percent of the food grains. The pattern of rainfall is continuously changing over time due to environmental disturbances. Weather forecasting is a difficult task and scientists have been working is this field to attain accuracy at maximum level. ARMA model are the time series-based methods to analyse and model the data values. The stationary behaviour of the time series is studied in the paper. Studied the Fractal and wavelet methods [1]. Studied T-170 model for forecasting [2]. Doppler weather radar data used for nowcast [3]. For forecasting temperature artificial intelligence technique were used [4]. Day to day average wind energy is forecasted using ANN [5]. Box-Jenkins method are used for analysing the environmental data[6]. ARIMA models are studied and best suited model is selected on the basis of the AIC value.

2. METHODOLOGY

A. Stationarity

If statistical values of a series such as mean, variance, covariance, autocorrelation are all independent of time. Stationarity is the most important and required condition for the application of AR., MA, ARMA, ARIMA models for studying and analysing and modelling the data. It is also important because it helps in identifying the driving factors such as trend and seasonality. Hence for the application of ARIMA model it is important for the series to be consistent.

B. Augmented-Dickey-Fuller (ADF)

ADF checks stationary of time series of given time period. ADF tests null hypothesis presence in the data values based on time. Null hypothesis (η_0) in the ADF test is that the data values based on time is non-stationary when the probability value is > 5% and hence (η_0) is accepted and it the probability value is < 5% then (η_0) is dismissed and alternative hypothesis (η_1) is accepted hence the series is stationary. The testing process of ADF is:

$$\Delta y_t = \phi + \varphi t + \chi y_{t-1} + \lambda_t \Delta y_{t-1} + \ldots\ldots\ldots + \ldots\ldots$$

where ϕ is the constant and φ the coefficient on the time trend

C. Philips Peron (PP)

PP test is used to check the stationarity. The Philips Peron test is mode to check serial correlation with application of Newey West (1987) heteroscedascity, autocorrelation consistent covariance matric analyser. The η_0 in the PP test is that the series is non-stationary when the probability value is > 5% and hence X_0 is accepted and it the probability value is < 5% then the η_0 is rejected and η_1 is accepted implying that the series is stationary. PP test helps us to detect stationarity in the time series.
Philips Peron equation can be seen as:

$$\Delta x_t = \alpha X_t + \pi x_{t-1} + a_t$$

D. Auto Regressive

Autoregressive AR model is used for the statistical calculations of a time series; hence estimating future values on the basis of the weighted sum of the previous values. The model shows that the resulting values linearly depends on the stochastic terms, hence the model is represented as the stochastic difference equation. AR (1) is referred to as the process of first order, which means that present value is calculated on the basis of the preceding value. AR(l) at level 'l'; which is:

$$\alpha_i = a + \sum_{t=1}^{l} \phi_t \alpha_{i-t} + \beta_i$$

- *Rashmi Bhardwaj*, Professor of Mathematics, University School of Basic & Applied Sciences (USBAS), Head, Non-Linear Dynamics Research Lab, Guru Gobind Singh Indraprastha University, Delhi, India, Email: rashmib22@gmail.com*
- *Varsha Duhoon, Research Scholar(s), USBAS, GGS Indraprastha University, Delhi, India, Email: varshaduhoon5@gmail.com*

β_i is the noise; a is constant term and $\phi = (\phi_1, \phi_2 \phi_3 \phi_4 \phi_5, ..., \phi_j)$ is the vector of model coefficients and 'j' is the positive Integer.

X is the lag operator and is the symbol of lagged values of the process, hence

$$X\omega_i = \omega_{i-1}, X^2\omega_i = \omega_{i-2}, X^3\omega_i = \omega_{i-3}, ..., X^d\omega_i = \omega_{i-d}.$$

Further,

$$\delta(X) = 1 - \sum_{i=1}^{N} \psi_j X^i = 1 - \psi_1 X - \psi_2 X^2 - ..... - \psi_N X^N$$

AR(l) process is expressed as:

$$\psi(X) Y_r = \lambda_r; r = 1, ..., n$$

$\psi(X)$ is characteristic polynomial; processing further its roots or factors or zeros determine the stationarity of the process.

E. Moving Average

Moving average model models single time series. It shows; resulting variables depends linearly on present and previous values of a stochastic term. It when combined with Auto regressive model forms the ARMA model that is Auto regressive Moving average model hence having more complicated stochastic structure. The process of order 'm' also written as MA(m) is:

$$X_t = \mu + \in_t + \sum_{i=1}^{m} \theta_i \in_{t-i}$$

where θ_i =parameters; $\mu = expectation\ of\ X_t$; $\epsilon_t \ldots . \epsilon_{t-i}$ = White Noise Error terms (is a random variable).

In the model, the process depends on past $\in_t$'s. MA(m) defines the correlated disturbance structure in our time series. Lag operator in MA(m) process is given as $A_i = \mu(X) \in_t$ where, $\mu(X) = 1 + \sum_{i=1}^{n} \gamma_i X^i$

.

F. AkaikeInformationCriterion (AIC)

It is used for analysing quality of statistical models applied on time series values. AIC is the method to analyse the best fitting model from the calculated models of AR, MA, ARMA, ARIMA. Thus, AIC estimates the quality of the models with respect to the other models, hence, AIC is way or method to select the best suited model among others. The calculation of AIC is based on statistical model.

$$AIC = 2k - 2InL$$

- k = number of calculated variables
- L = higher value of the likelihood function

G. Auto correlation Function (ACF)

Auto correlation can be defined as the correlation between the signals with the delayed copy of itself. ACF is the method to study the similarity among the values which are function of time gap among them. ACF is the mathematical tool to analyse and track repeating patterns which can be detecting the missing values in the signal and existence of periodic signals obscured by noise.

$$\rho = \frac{\mathrm{cov}(y_1, y_2)}{\sigma_1 \sigma_2}$$

- y_1, y_2, = Variables;
- $\sigma_1 \sigma_2$ = Standard Deviation.

H. Partial Auto Correlation Function (PACF)

PACF is the conditional correlation, which is the study of the relationship between two variables.

$$y_t = \phi_{21} y_{t-1} + \phi_{22} y_{t-2} + \in_t$$

- y_t = Original series minus the Sample Mean,
- $\emptyset_{ij}$ = Value of Partial Auto Correlator of a particular order.

I. Auto Regressive Moving Average Model (ARMA)

To model data and further forecast AR and MA models are combined to produce ARMA. AR represents regressing the variable with the past values. MA represents the modelling of error values in the form of linear combinations.

$$X_t = c + \in_t + \sum_{n=1}^{m} \theta_n \in_{t-n} + \sum_{n=1}^{l} \theta_n k_{t-n}$$

ARMA is symbolised as (l, m); l autoregressive model and m moving average model. Lag operator in ARMA (l, m):

$$\phi(X) A_i = \theta(X) A_i$$

J. Auto Regressive Integrated Moving Average Model (ARIMA)

This is applied to the data to analyse, model and forecast the time series value. AR in ARIMA implies to the regression of the variable on its own lagged values. The MA part of ARIMA refers to the linear combination of the regression errors. I in ARIMA refers to integrated which means that the time-based values are substituted by gap of the present values and past values. ARIMA model is also denoted by (l, d, m); the values of l, d, m are non-negative that is they are positive. Now, l is number of time lags, d: degree of gap and m: order.

ARIMA:

$$(1 - \sum_{t=1}^{l} \alpha_i L^t)(1 - L)^d X_i = (1 - \sum_{t=1}^{m} \alpha_t L^t) \varepsilon_t$$

ARMA is represented as follows:

$$X_t = c + \in_t + \sum_{h=1}^{m} \theta_i \in_{t-h} + \sum_{h=1}^{l} \theta_i k_{t-h}$$

Now, extending ARMA model in order to add differencing, the following steps are taken:

$$A_k = Z_k - Z_{k-1}$$

$$A_k - A_{k-1} = Z_k - 2Z_{k-1} + Z_{k-2}$$

$$A_k - \sum_{j=1}^{d} A_{k-j} = (1 - X)^d Z_k$$

Where, 'd' is the order of differencing. Further, substituting $\in_t$ in ARMA model, yields the formula:

$$(1 - \sum_{t=1}^{l} \alpha_i L^t)(1 - L)^d X_i = (1 - \sum_{t=1}^{m} \alpha_t L^t) \varepsilon_t$$

3. RESULTS & DISCUSSION

Daily data of rainfall for Delhi with coordinates Longitude 77^0 09' 27'' Latitude 28^0 38 '23'' N Altitude :228.61m has been taken from January 1, 2017 up till December 31, 2018. The time series of daily values of rainfall have been taken and the ADF and PP test are applied so as to see whether the series shows Stationarity or not on the basis of the probability values. Later on, on the basis of the AIC values the best suited model of ARIMA is selected for forecasting the 7 days values of rainfall on daily basis.

η_0 : Rainfall has unit root. (Non-Stationary) and η_1 : Rainfall does not have unit root. (Stationary)

Table I: *The probability value of ADF and PP test*

Test	Probability value	Accept/ Reject η_0 Hypothesis	Stationary/Non-Stationary
ADF TEST	0.000 (<5%)	Reject η_0	Stationary
PP TEST	0.000 (<5%)	Reject η_1	Stationary

Hence, since the time series has now, proved to be the stationary now, further apply different ARIMA models in or to choose best model among all on basis of AIC value of each model.

Table II: *Models and AIC vales for each model for rainfall*

MODEL	(l, d, m)	AIC VALUE
AR	(1,0,0)	5638.154
MA	(0,0,1)	5640.766
ARMA	(1,0,1)	5589.394
ARIMA	(1,1,1)	5587.265

In Table II, ARIMA (1, 1,1) model shows the least value of AIC and hence the further forecasting of time series of rainfall is done using ARIMA (1, 1, 1) model. Now, the PACF and ACF graphs are as follows:

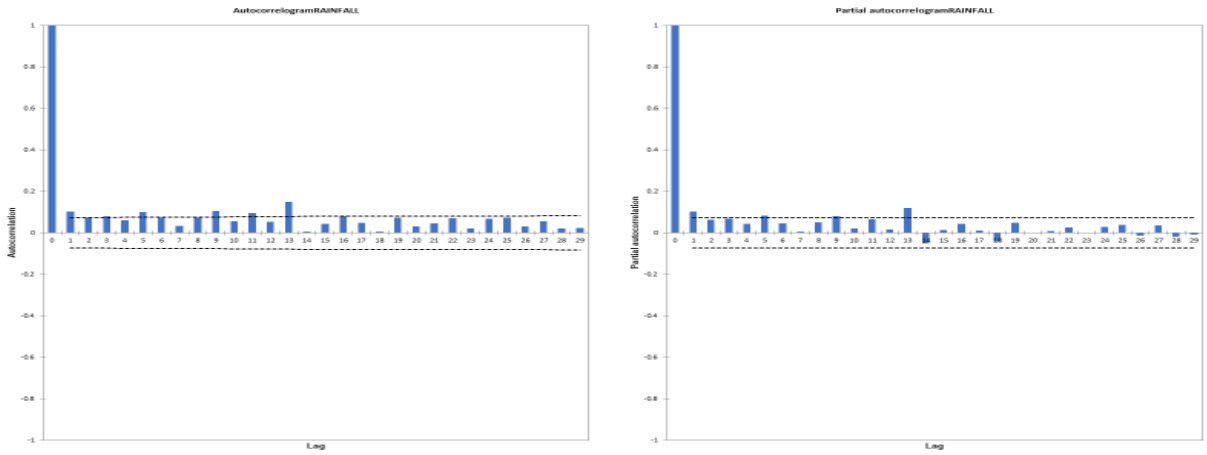

Fig. 1: *Partial Auto Correlogram, Auto Correlogram Of Rainfall ARIMA (1,1,1) model*

Above fig. 1, shows the ACF and PACF values of the rainfall data obtained using the ARIMA model. Now, below is the residual plot of rainfall time series in fig. 2.

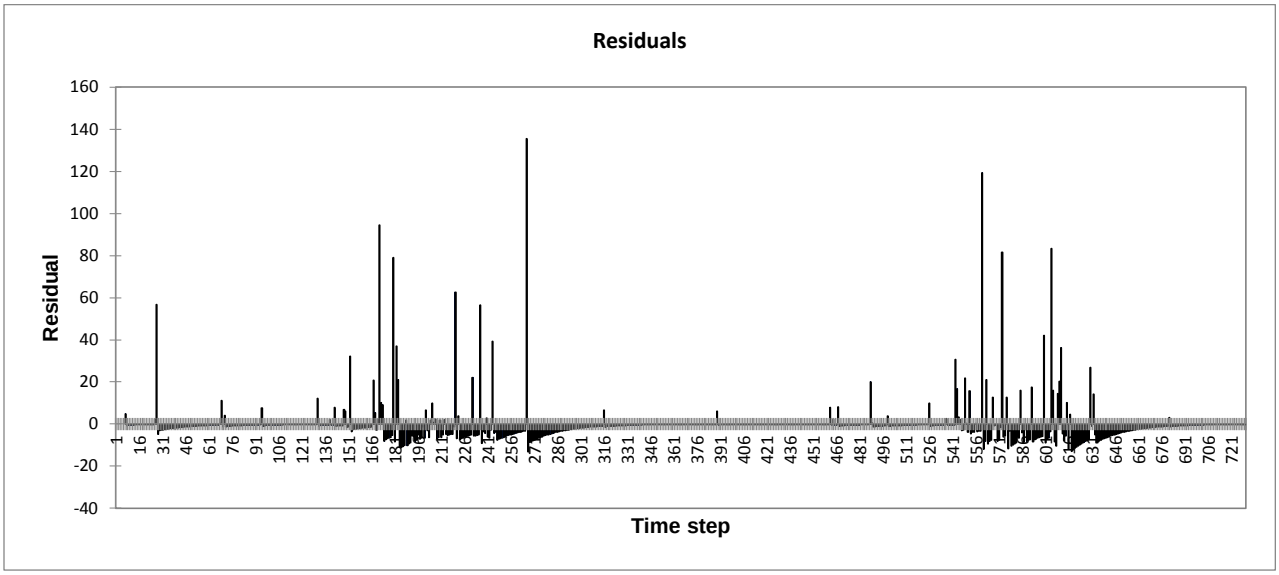

***Fig. 2**: Residual plot of Rainfall ARIMA (1,1,1) model*

Fig. 2 shows the residual plots of the time series using the ARIMA (1,1,1) model at equal time steps. The residual is least in the ARIMA (1,1,1) model. Now, the actual and predicted values using the ARIMA (1,1,1) has been done.

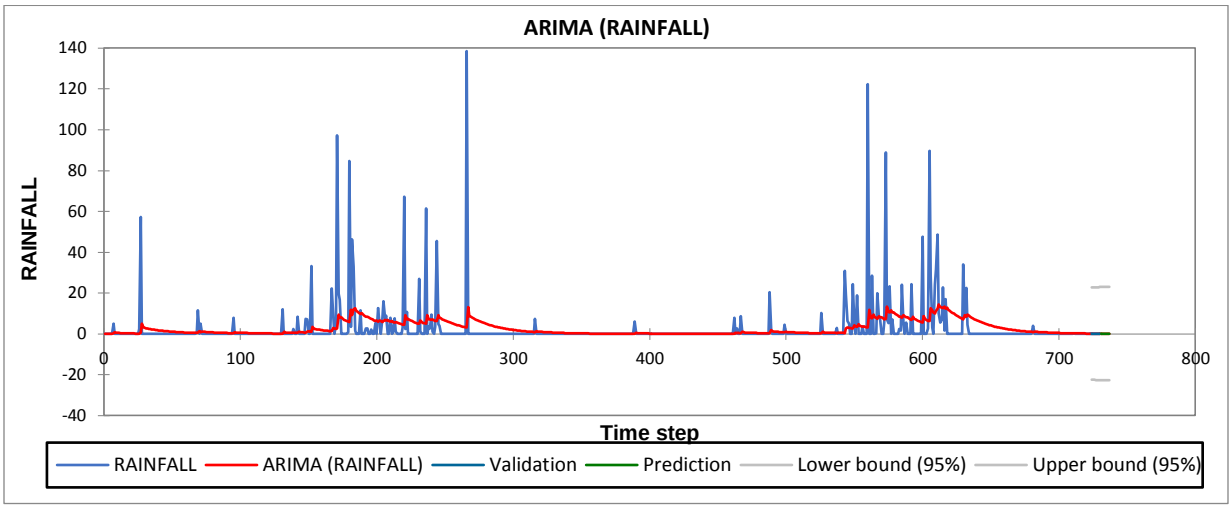

***Fig. 3:** Actual and Forecasted values of Rainfall using ARIMA (1,1,1) model*

In fig. 3, the blue line shows the actual data of rainfall, the red line shows ARIMA values of rainfall, green line shows the predicted values of rainfall for further using ARIMA.

4. CONCLUSION

The stationary behaviour of rainfall pattern is studied using ARIMA model. ARIMA (1,1,1) model proved to be the best suited model in order to forecast daily rainfall for Delhi region which was selected on the basis of the AIC value among the AR (1), MA (1), ARMA (1,1) and ARIMA (1,1,1) models. Also, ADF and PP test clearly showed the rainfall data values are consistent which means statistical measures of the series is not dependent of time.

ACKNOWLEDGEMENT

The authors are thankful to Guru Gobind Singh Indraprastha University, Delhi(India)for providing researchfacilities and financialsupport.

REFERNCES

[1] Bhardwaj R. “Wavelets and Fractal Methods with environmental applications.” Mathematical Models, Methods and Applications. Eds. Siddiqi, Manchanda, Bhardwaj, 2016, 173-195.

[2] Bhardwaj R., Kumar A., Maini P., Kar S.C., Rathore L.S. “Bias free rainfall forecast and temperature trend-based temperature forecast based upon T-170 Model during monsoon season”. Meteorological Applications. 2007, 14(4), 351-360.

[3] Bhardwaj R., Srivastava K. “Real time Nowcast of a Cloudburst and a Thunderstorm event with assimilation of Doppler Weather Radar data.” Natural Hazards. 2014, 70(2), 1357-1383.

[4] Bhardwaj R., Duhoon V., "Weather forcasting using Soft Computing Technique", International Conference on Computing, Power, and Communication Technologies (GUCON), IEEE Explore Digital Library. 2018, pp1111-1115.

[5] Dumitru C. D., Gligor A., 2017, "Daily average wind energy forecasting using ANN", Procedia Engineering. 181, 2017, pp 829-836.

[6] Mihai, M., Meghea I., 2011, "Box Jenkins methodology applied to the environmental monitoring data", Applied Sciences. 13, 2011, pp. 74-81.