Forecasting Rainfall Pattern in Western Uttar Pradesh, India: A time series analysis using ARIMA

Abin Johnson, Apparasu Sruthi

Abstract

The prospect of adequate rainfall is a matter of concern for agricultural and environmental practitioners and policy makers. In this context, time series modeling and forecasting is being used to support these applications in forecasting rainfall patterns. The aim of this study is to make use of a well-used time series model ARIMA in forecasting annual rainfall in Uttar Pradesh (West) sub division, by making use of historical data of over 100 years (from 1901-2015). The data thus collected is transformed into a suitable model using R software, and it was found that ARIMA (0,1,3) is a model suitable for the given data set. As such, this model was used to forecast the pattern of annual rainfall in the region for twenty years (2016-2035) – which can be used in further research and policy making purposes. Also, the study theoretically notes the concepts of AR, MA, ARMA and SARIMA – where and how do we use them

Keywords: Rainfall in India, Uttar Pradesh, Time Series Data, ARIMA, forecasting

Introduction

The prediction of the future courses of meteorological quantities on the basis of historical time series is a significant base for several fields – especially agricultural production and the related policy dynamics. This is especially true for rainfall as the crop production is highly determined by the relative changes in it. Obvious climatic changes aside, rainfall as an input in productivity can be forecasted – rather easily – by the time series models that make use of historical data, of the past patterns in that data to forecast the future occurrences.

Time series is observation of a variable at discrete points of time (usually equals spaced) that is measured and sorted according to time (Chatfield, 2001). This technique is used to explain data using statistical and graphical methods, to select the best statistical models to explain the data generating process, to predict the future amounts of a series and controlling a given process (Radhakrishnan and Dinesh, 2006). Several methods are used in rainfall forecasting using time series– including regression analysis, exponential smoothing and auto-regressive integrated moving average (ARIMA). This paper especially deals with ARIMA – one of the well-known linear models for time series modeling and predicting (Mirzavand and Ghazavi, 2015). It is a synthesis of the Autoregressive and Moving Average models; the details of which will be dealt in detail.

Several other models are also available for rainfall forecasting. Selecting a suitable technique for modeling a phenomenon depends on various factors such as data accuracy, time, cost, ease of use of the model's results, interpretation of results, etc. (Mondal and Wasimi, 2007). Several researches have already supported and were successful in predicting various environmental and weather-related parameters by using ARIMA model.

Theoretical Framework

Generally, the models for time series data can have different forms and represent different non-deterministic processes (Sokolnikov, 2013). Most modeling of time series like AR, MA and ARMA models have linear bases (Mirzavand & Ghazavi, 2015). In this research, we make use of ARIMA – a synthesis of AR and MA with an integrating difference factor.

Autoregressive Model (AR): In an autoregression model, we forecast the variable of interest using a linear combination of *past values of the variable*. The term *auto*regression indicates that it is a regression of the variable against itself.

The equation for an AR model of order p is:

$$Yt = \delta + \varphi 1Yt - 1 + \varphi 2Yt - 2 + \varphi 3Yt - 3 + \cdots ..+\varphi pYt - p + \varepsilon p$$

Moving Average (MA): Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting. They are the covariance stationary that can be used for a wide variety of autocorrelation patterns. Mean of an MA process is constantly zero.

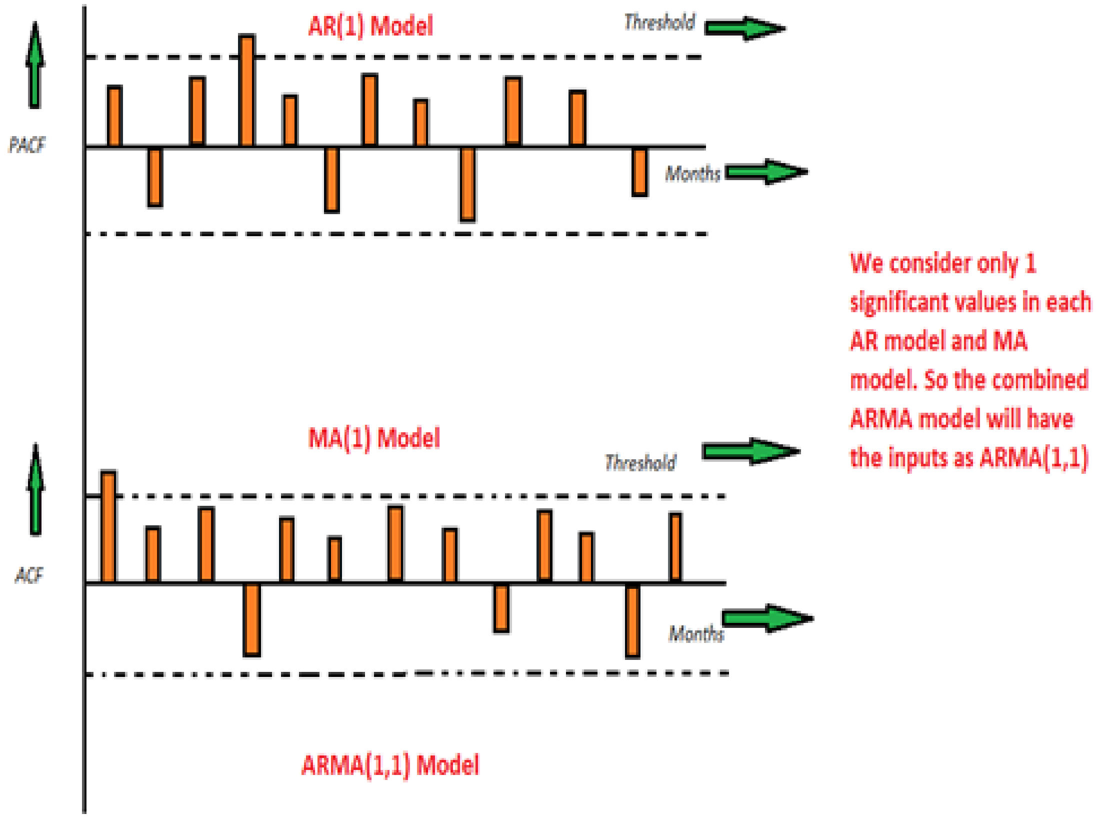
Moving Average model can be expressed as:

$$Yt = \alpha_1 * \mathcal{E}_t -{}_1 + \alpha_2 * \mathcal{E}_t -{}_2 + \alpha_3 * \mathcal{E}_t -{}_3 + \ldots\ldots\ldots\ldots + \alpha_k * \mathcal{E}_t -{}_k$$

Autoregressive Moving Average (ARMA): This is a model that is combined from the AR and MA models. ARMA models form a type of linear models which are widely applicable and parsimonious in parameterization.

Here β represents the coefficients of the AR model and α represents the coefficients of the MA model.

$$Yt = \beta_1 * y_{t-1} + \alpha_1 * \mathcal{E}_{t-1} + \beta_2 * y_{t-2} + \alpha_2 * \mathcal{E}_{t-2} + \beta_3 * y_{t-3} + \alpha_3 * \mathcal{E}_{t-3} + \ldots + \beta_k * y_{t-k} + \alpha_k * \mathcal{E}_{t-k}$$

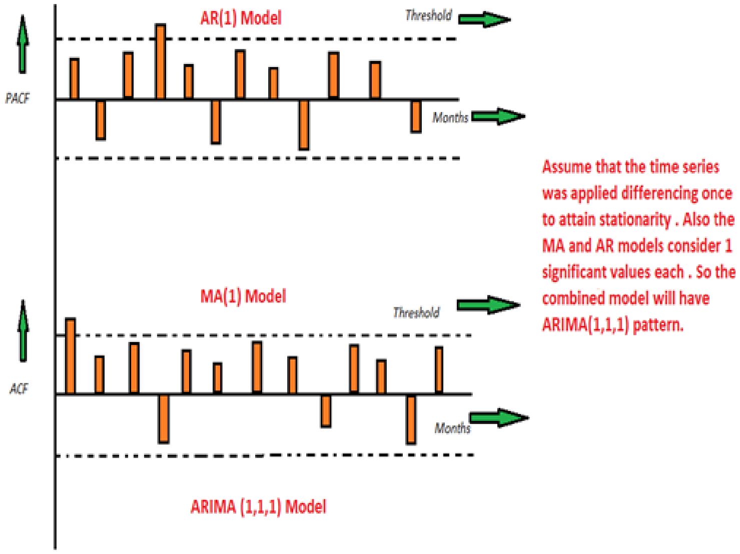


We consider only 1 significant values in each AR model and MA model. So the combined ARMA model will have the inputs as ARMA(1,1)

Autoregressive Integrated Moving Average (ARIMA): ARIMA model is a generalization of a simple autoregressive moving average.

The ARIMA model is quite similar to the ARMA model other than the fact that it includes one more factor known as integrated differencing step, i.e., removing the seasonal and trend components. ARIMA models have originated from the synthesis of AR and MA models. ARIMA model is generally denoted as ARIMA(p, d, q) and parameter p, d, q are defined as follow:

- p: the lag order or the number of time lag of autoregressive model AR(p)
- d: degree of differencing or the number of times the data have had subtracted with past value
- q: the order of moving average model MA(q)


AR(1) Model
Threshold
PACF
Months
Assume that the time series was applied differencing once to attain stationarity . Also the MA and AR models consider 1 significant values each . So the combined model will have ARIMA(1,1,1) pattern.
MA(1) Model
Threshold
ACF
Months
ARIMA (1,1,1) Model

Seasonal Autoregressive Integrated Moving Average (SARIMA): In ARIMA, the future amount of a parameter is assumed to be a linear function of past observations and random errors. A SARIMA model can be explained as ARIMA (p, d, q) (P, D, Q)s, where (p, d, q) is the non-seasonal component of the model and (P, D, Q)s is the seasonal component of the model.

Here, p is the order of non-seasonal autoregressive, d is the number of regular differencing, q is the order of nonseasonal Moving Average. Also, P is the order of seasonal autoregression, D is the number of seasonal differencing, Q is the order of seasonal Moving Average, and s is the length of season. Thus, by introducing three more hyperparameters, SARIMA encompasses in itself both trend and seasonality components.

Review of Literature

As mentioned earlier, the techniques of time series modeling, especially ARIMA, was extensively used in research related to weather forecasting.

In his bestselling work “Time Series Forecasting” (2001), Chatfield defined time series as a set of observations measured sequentially through time. He said that description, modeling, forecasting and control are the objectives of any time series analysis – the process we follow in the course of this paper. Describing time series as a similar process, Radhakrishnan & Dinesh (2006) implemented an alternative approach of barcode scan to analyze rainfall behavior in Malaysia.

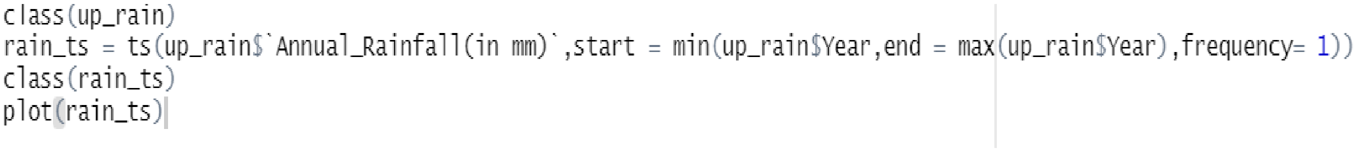
A study by Hazarika, et.al. (2016) to forecast the pattern of monthly rainfall in Assam used the Box-Jenkins method of ARIMA and SARIMA in three phases in R software– data preparation and model selection, estimation and diagnosis, followed by forecasting. Our study follows a similar pattern too. Murat, et.al. (2018) also used ARIMA to get sensible forecasts of meteorological data daily air temperature and precipitation in R.

Several studies, however, did a comparative analysis of the time series models to find a best one for forecasting the future behavior of variables. Of them were researches by Mirzavand & Ghazvi (2015) who found AR model best for predicting groundwater level in Iran; who concluded that combining time series models have an advantage in terms of groundwater level forecasting. Dastorani et. al (2016) also assessed several models to forecast monthly rainfall, only to find MA model showed the best performance for the data (33%). Besides, Bang et.al (2019) predicted crop yield in India by means of temperature and rainfall parameters using ARMA, SARIMA and ARMAX models.

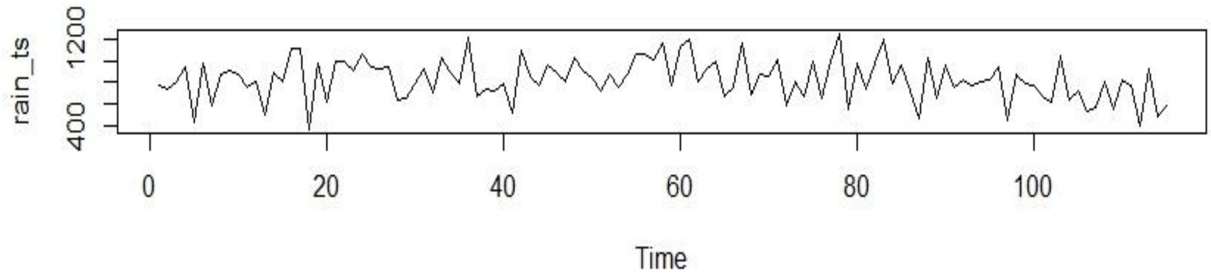
Data Analysis

For this study, the time series data of the annual rainfall (in mm) in Uttar Pradesh (West) sub-division from 1901-2015 is sourced and extracted from the Indian Meteorological Department. The huge dataset consisted of monthly, seasonal and annual rainfall figures of all the divisions in the country in the said time-period.

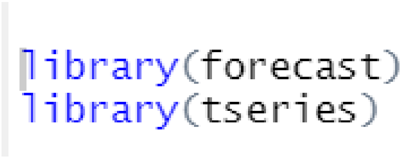
After importing the data into R, the first step in the process is converting the available data into time series format using $ts()$ function, which is used to create a time-series object. Since it is an annual data starting from 1901, we set the frequency as 1. When checked using $class$ and $plot$ functions, the time series data was confirmed for further analysis.



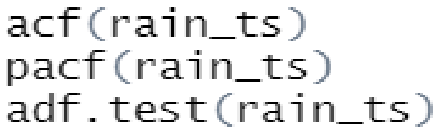
```
class(up_rain)
rain_ts = ts(up_rain$`Annual_Rainfall(in mm)`,start = min(up_rain$Year,end = max(up_rain$Year),frequency= 1))
class(rain_ts)
plot(rain_ts)
```



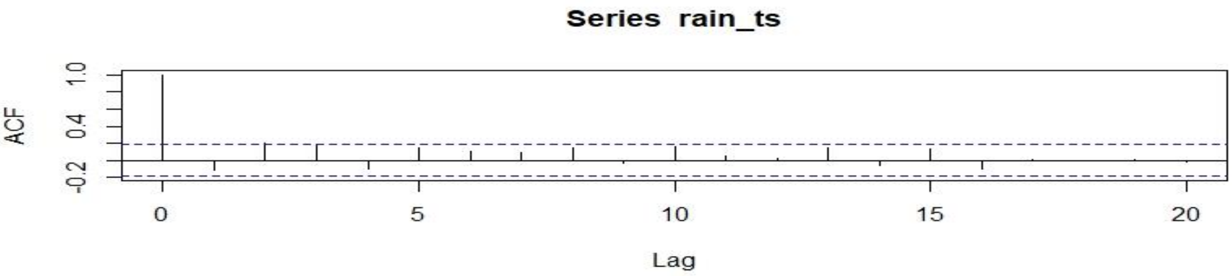
Two packages namely $forecast$ and $tseries$ were imported into R which provides methods and tools for displaying and analyzing univariate time series forecasts including exponential smoothing via state space models and automatic ARIMA modeling. $tseries$ is especially used in computational finance. Then, we run the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots on the obtained time series data. Then, we also check for stationarity of the series using the augmented Dickey-Fuller test $(adf.test)$.

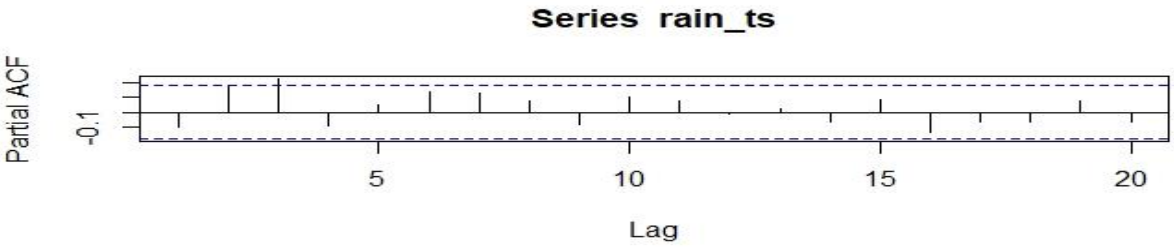


```
library(forecast)
library(tseries)
```

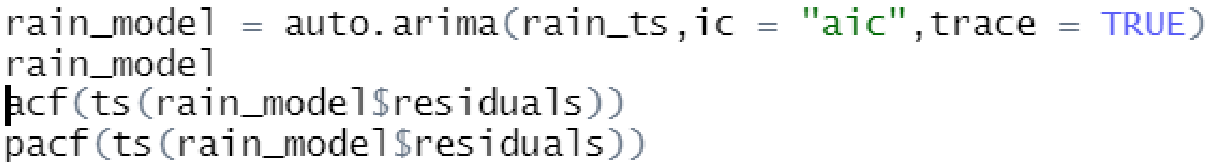


```
acf(rain_ts)
pacf(rain_ts)
adf.test(rain_ts)
```

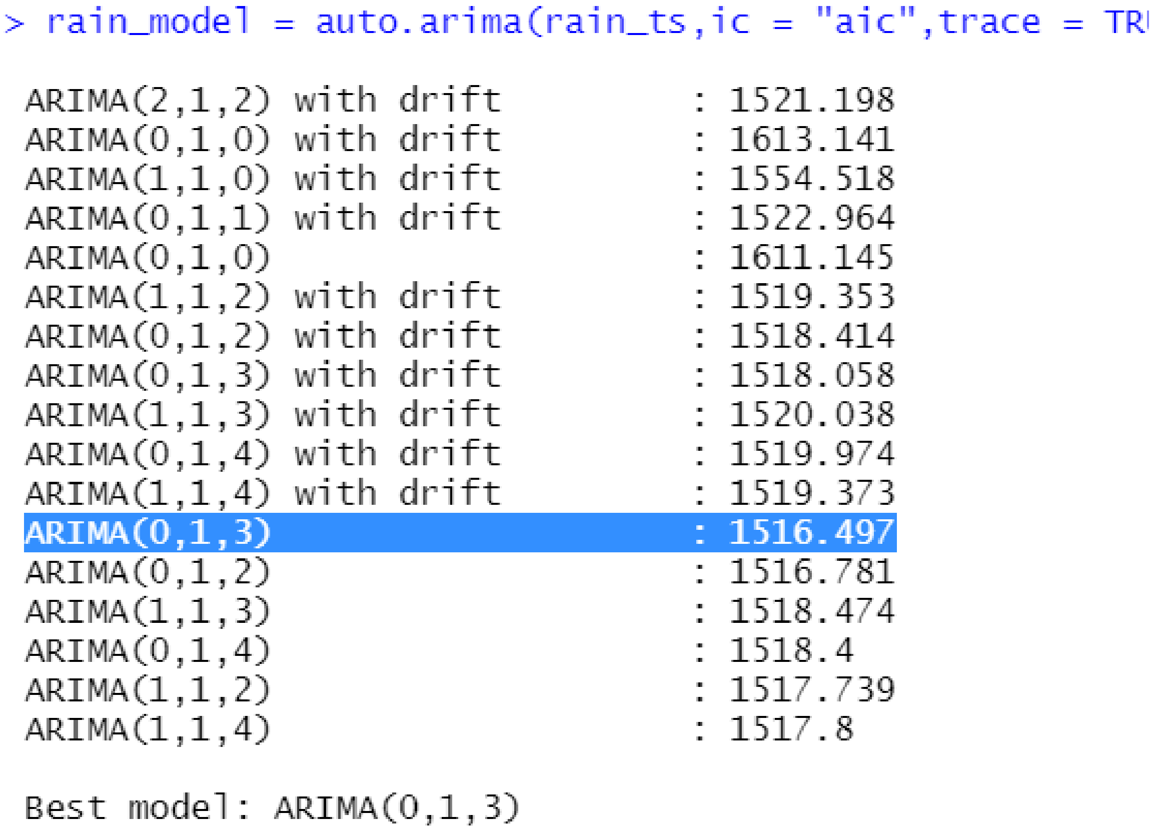




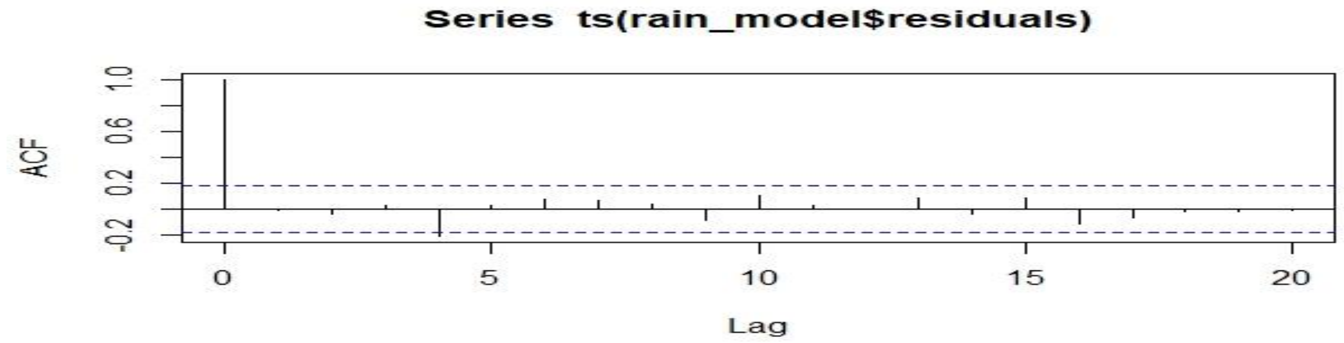
Later, $auto.arima$ function is used – that combines unit root tests, minimisation of the Akaike Information Criterion (AIC) and Maximum Likelihood Estimation (MLE) – to obtain a best ARIMA model. This best model is then used to fit the processed time series data. The ACF and PACF plots of the fitted model can be obtained.

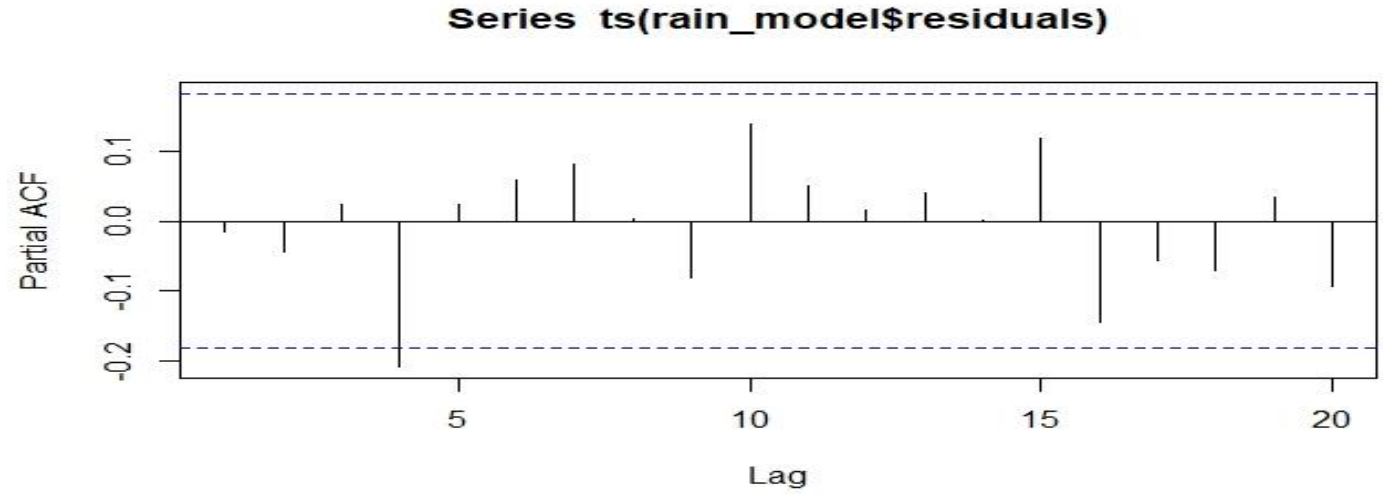


```
rain_model = auto.arima(rain_ts,ic = "aic",trace = TRUE)
rain_model
acf(ts(rain_model$residuals))
pacf(ts(rain_model$residuals))
```

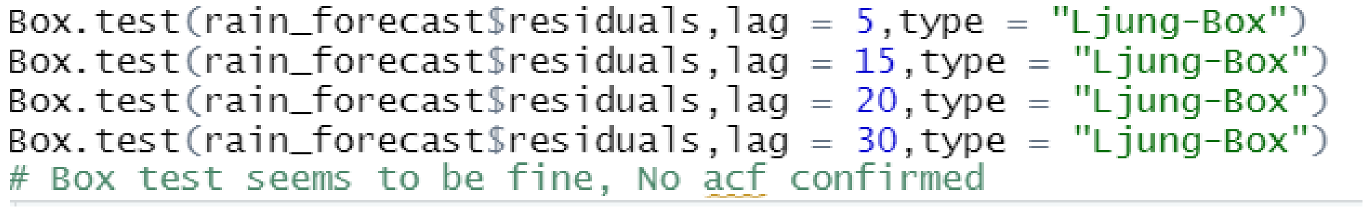


```
 > rain_model = auto.arima(rain_ts,ic = "aic",trace = TR

   ARIMA(2,1,2) with drift         : 1521.198
   ARIMA(0,1,0) with drift         : 1613.141
   ARIMA(1,1,0) with drift         : 1554.518
   ARIMA(0,1,1) with drift         : 1522.964
   ARIMA(0,1,0)                    : 1611.145
   ARIMA(1,1,2) with drift         : 1519.353
   ARIMA(0,1,2) with drift         : 1518.414
   ARIMA(0,1,3) with drift         : 1518.058
   ARIMA(1,1,3) with drift         : 1520.038
   ARIMA(0,1,4) with drift         : 1519.974
   ARIMA(1,1,4) with drift         : 1519.373
   ARIMA(0,1,3)                    : 1516.497
   ARIMA(0,1,2)                    : 1516.781
   ARIMA(1,1,3)                    : 1518.474
   ARIMA(0,1,4)                    : 1518.4
   ARIMA(1,1,2)                    : 1517.739
   ARIMA(1,1,4)                    : 1517.8

   Best model: ARIMA(0,1,3)
```

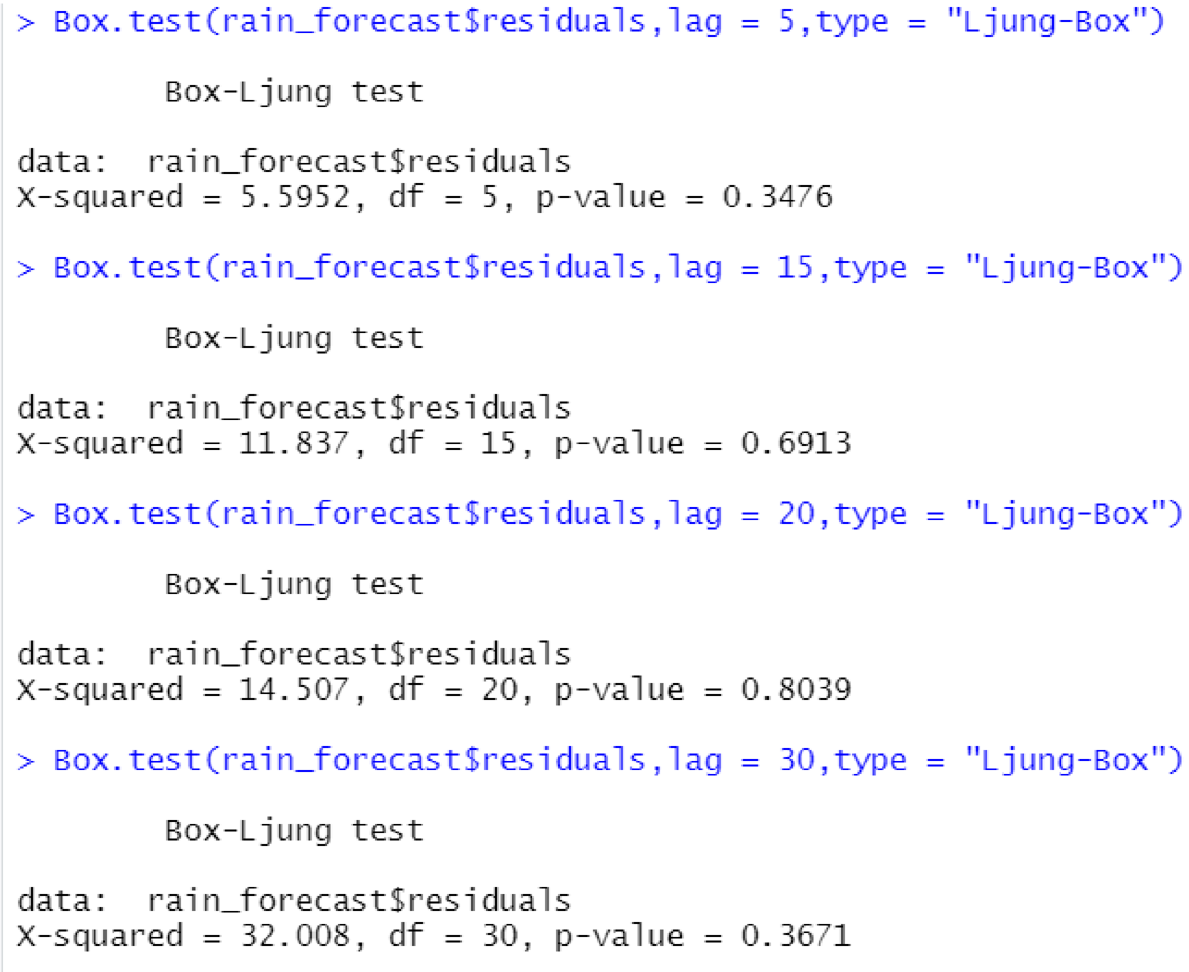




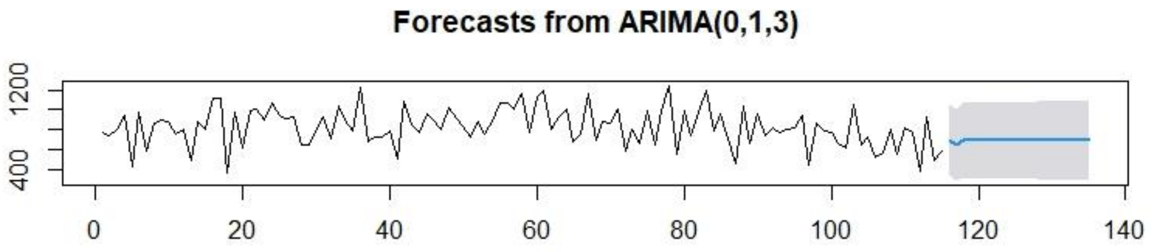
This ARIMA model of the obtained time series can be used for forecasting. For that, generic $forecast$ function is used with a confidence level of 95%, to forecast the rainfall variable for the next 20 years (in $level$, c is the confidence interval and h is the number of periods for forecasting). This forecast can be plotted using $plot$ function. Then, we compute the Box--Pierce or Ljung--Box test statistic using $box.test$ for examining the null hypothesis of independence in the time series. For different values of lag in the function, we validate our forecasts using the null and alternative hypothesis of the Box-Ljung test.



```
Box.test(rain_forecast$residuals,lag = 5,type = "Ljung-Box")
Box.test(rain_forecast$residuals,lag = 15,type = "Ljung-Box")
Box.test(rain_forecast$residuals,lag = 20,type = "Ljung-Box")
Box.test(rain_forecast$residuals,lag = 30,type = "Ljung-Box")
# Box test seems to be fine, No acf confirmed
```



```
> Box.test(rain_forecast$residuals,lag = 5,type = "Ljung-Box")

        Box-Ljung test

data:  rain_forecast$residuals
X-squared = 5.5952, df = 5, p-value = 0.3476

> Box.test(rain_forecast$residuals,lag = 15,type = "Ljung-Box")

        Box-Ljung test

data:  rain_forecast$residuals
X-squared = 11.837, df = 15, p-value = 0.6913

> Box.test(rain_forecast$residuals,lag = 20,type = "Ljung-Box")

        Box-Ljung test

data:  rain_forecast$residuals
X-squared = 14.507, df = 20, p-value = 0.8039

> Box.test(rain_forecast$residuals,lag = 30,type = "Ljung-Box")

        Box-Ljung test

data:  rain_forecast$residuals
X-squared = 32.008, df = 30, p-value = 0.3671
```



Results & Conclusion

The aim of this study was to forecast the pattern of rainfall in Uttar Pradesh (West) sub division from the historical data of over 100 years using the time series model of ARIMA. After the data was converted suitable for time series analysis, we checked for stationarity assumption using Dickey-Fuller test. Upon confirmation of stationarity in the data, we looked for a best fit model of ARIMA which was used to fit in our series. We then proceeded to forecast the rainfall pattern for the next 20 years and plotted it. Later, we validated our best-fit ARIMA model.

The forecasts thus obtained from ARIMA (0,1,3) model give us an idea of expected rainfall in the region during 2016-2035. However, the plot reveals a flat line after a certain point. This doesn't necessarily mean that the forecast is wrong – it indicates that there is no trend/ seasonality present in the data. Since the forecast of the future observation in the time series is conditional on the past observations, a flat line means that the forecast is showing mean rainfall. It is possible that we can force seasonality in the ARIMA model (theoretically: by making it a SARIMA) to get a better forecast for rain patterns.

Although the chosen model could not get the exact forecast for rainfall in Uttar Pradesh (West), it can give us information that helps to establish strategies for proper planning of agriculture or can be used as a supplemental tool for environmental planning and decision-making.

References

1. Bang, S., Bishnoi, R., Chauhan, AS., Dixit, AK. & Chawla, I. (2019). Fuzzy Logic based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA, and ARMAX models. *Twelfth International Conference on Contemporary Computing (IC3).*
2. Chatfield, C. (2001). Time Series Forecasting. *Boca Raton, Florida: Chapman & Hall/CRC Press.*
3. Dastroni, M., Mirzavand, M., Dastroni, MT. & Sadatinejad, SJ. (2016). Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate conditions. *Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards*.
4. Hazarika, B., Pathak, B. & Patowary, AN. (2016). Studying monthly rainfall over Dibrugarh, Assam: Use of SARIMA approach. *Mausam, Volume 68.*
5. Murat, M., Malinowska, I., Gos, M., Krzyszczak, J. (2018). Forecasting daily meteorological time series using ARIMA and regression models. *Int. Agrophys., 32(2), 253-264.*
6. Radhakrishnan, P. & Dinesh, S. (2006). An alternative approach to characterize time series data: Case study on Malaysian rainfall data. *Chaos, Solitons & Fractals, Volume 27, Issue 2.*