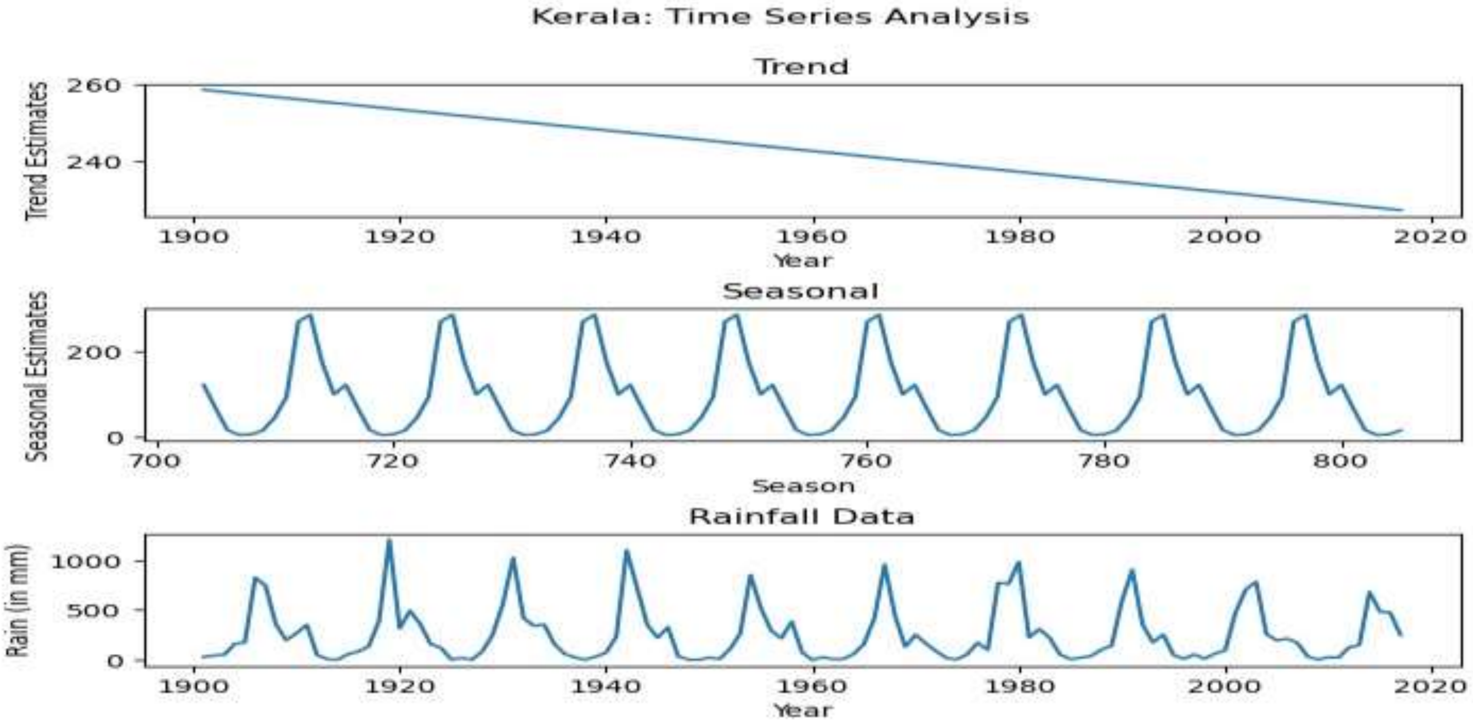
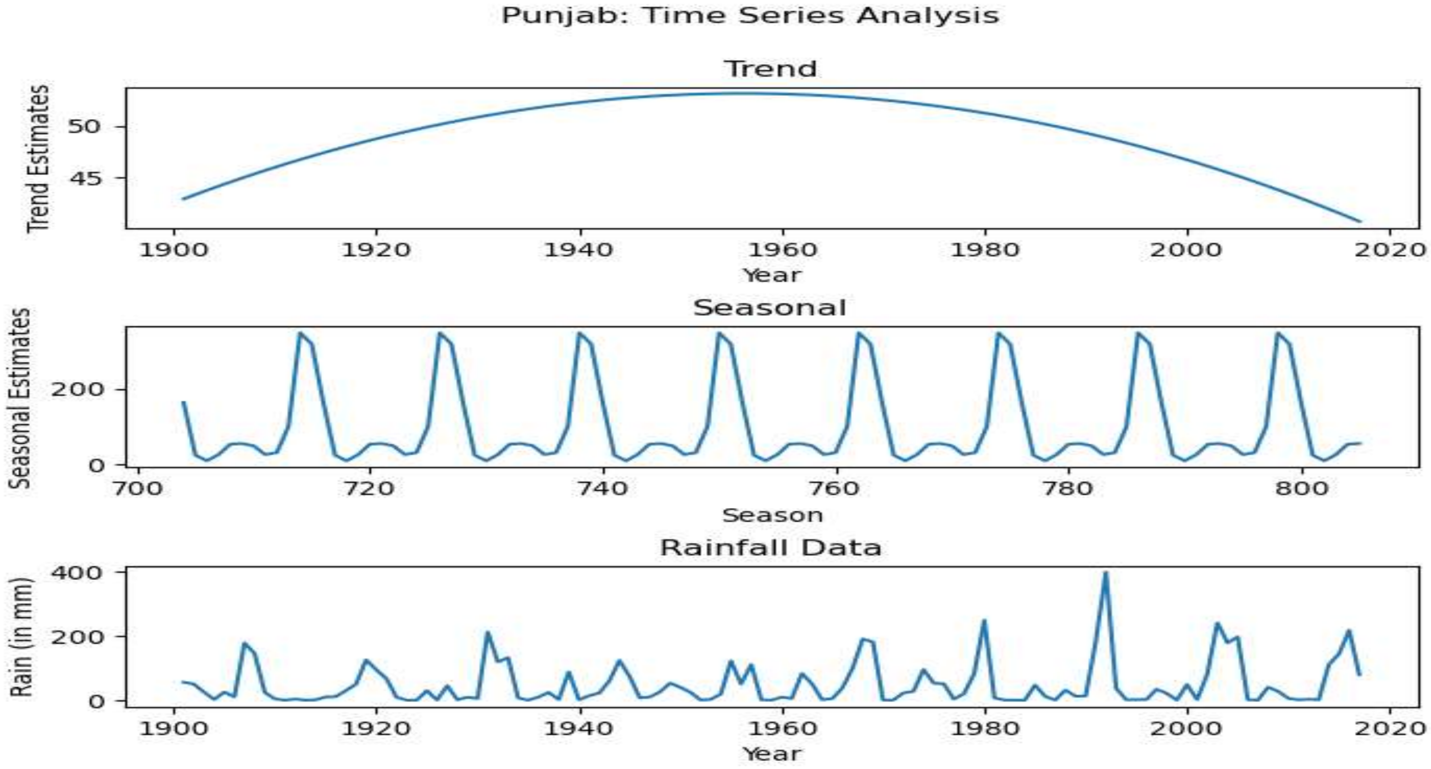
TIME SERIES ANALYSIS: ANALYSIS OF RAINFALL DATA IN INDIA (PART 2)

-Aastha Sumra, Kartikeya Sinha, Navya Garg, Nishika Taneja, Riya Khandelwal

ABSTRACT

The analysis of the rainfall data of the two states - Kerala and Punjab, for the years 1901-2017 showed that the rainfall data predominantly includes trend and seasonal pattern. We assumed, on the basis of the time plot and research, that the presence of cyclic component is negligible in the data. Using the Box-Jenkins approach, we fitted SARIMA models, for the years 1901-2000, on the basis of earlier analysis. The fitted models were judged on the basis of the predicted values for the years 2001-2017. Forecasts were obtained up to the year 2023.



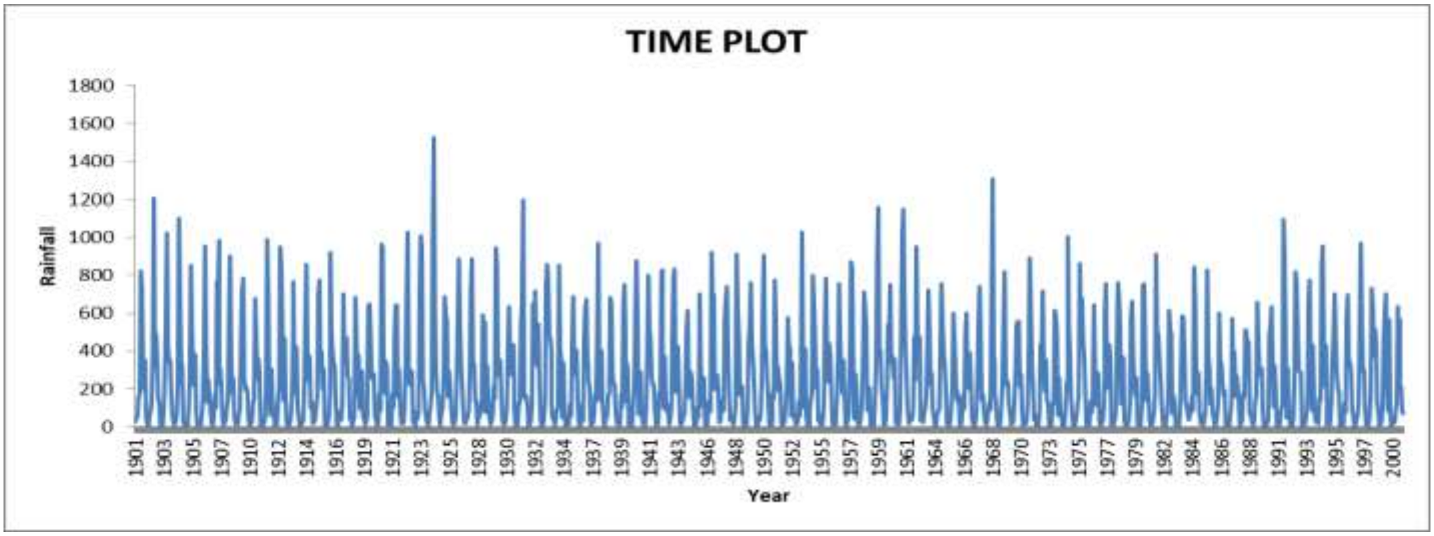


BOX JENKINS APPROACH TO FORECASTING

The Box Jenkins approach to forecasting was applied to both the states, Kerala and Punjab in R. We have divided our data into 2 parts-train data (years 1901-2000) and test data (years 2001-2017). We have carried out the fitting on the train data, and evaluated our model on the test data. Forecasting was carried out for the years 2018-2023.

1) KERELA

Stationarity



One of the basic assumptions of stationarity of the data is clearly visible from our time plot. We confirmed this by Augmented Dickey-Fuller test at 5% level of significance.

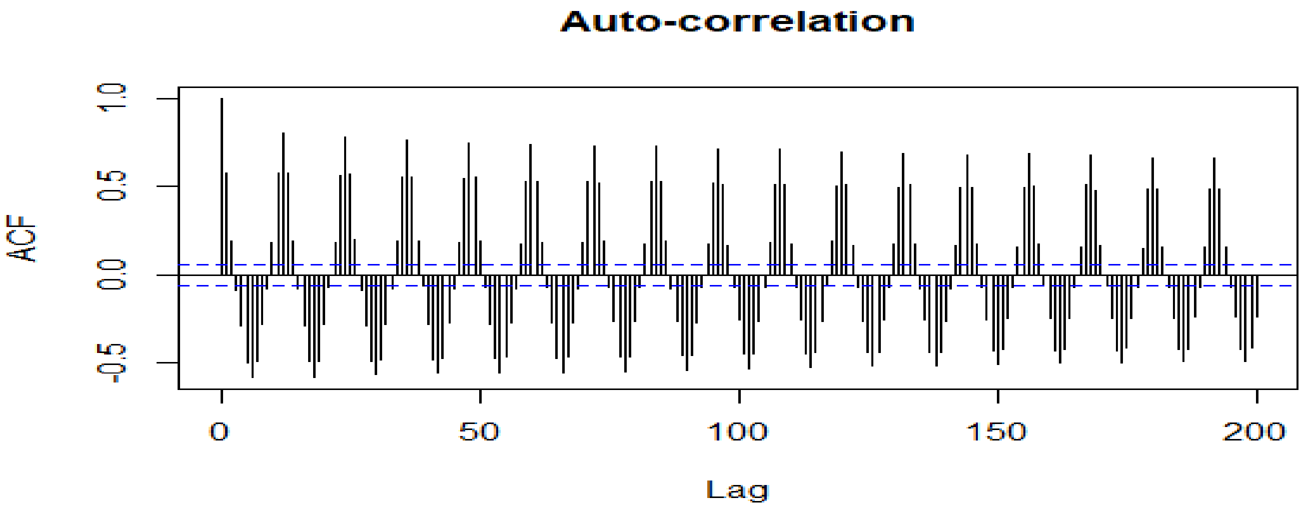
H_0 : The time series is non-stationary
H_1 : The time series is stationary

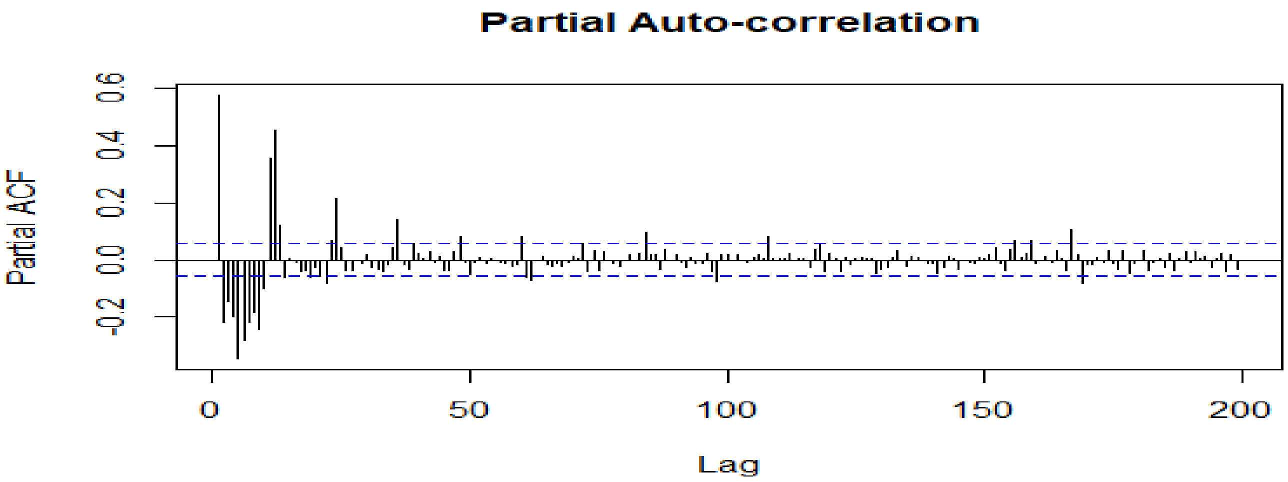
Level of significance = 5%
Result: Dickey-Fuller = -13.518, Lag order = 10, p-value = 0.01

Since, p-value < 0.05, therefore, we reject the null hypothesis. Hence, our **assumption of stationarity is correct**.

AC.F and PAC.F

The ac.f and pac.f were plotted to identify the relationships within the time series data and to check which type of model i.e., AR/MA/ARMA/ARIMA/SARIMA will be best suitable for our data. The ac.f plot shows the correlation of the series with its own lagged values and pac.f plot shows the correlation between the series and its lagged values after removing the correlations explained by intervening lags. The significant spikes in the ac.f and pac.f plots beyond the confidence interval guide us in the selection of the best model.





From the ac.f plot, it is visible that there are significant spikes in the plot at lags with period=12. So, there is a seasonal component present which should be accounted while choosing a model for our data. Since, there are significant spikes in both ac.f and pac.f plots , so an ARIMA model with seasonal parameters i.e., **SARIMA model with no. of seasons = 12** will be the best suitable model for our data.

Model Fitting

We fitted a SARIMA model on our data using auto.arima() function which gives us the model as:

ARIMA(3,0,2)(2,1,0)[12] with drift

```
Coefficients:

          ar1      ar2        ar3      ma1     ma2       sar1       sar2      drift

      -0.3255  -0.5178  -0.0377  0.3398  0.4996  -0.6197  -0.3121  -0.0412

s.e.    0.7779   0.4669   0.0334  0.7776  0.4958   0.0286   0.0283   0.1670

sigma^2 = 18511:  log likelihood = -7521.16

AIC=15060.31   AICc=15060.46   BIC=15106.03
```

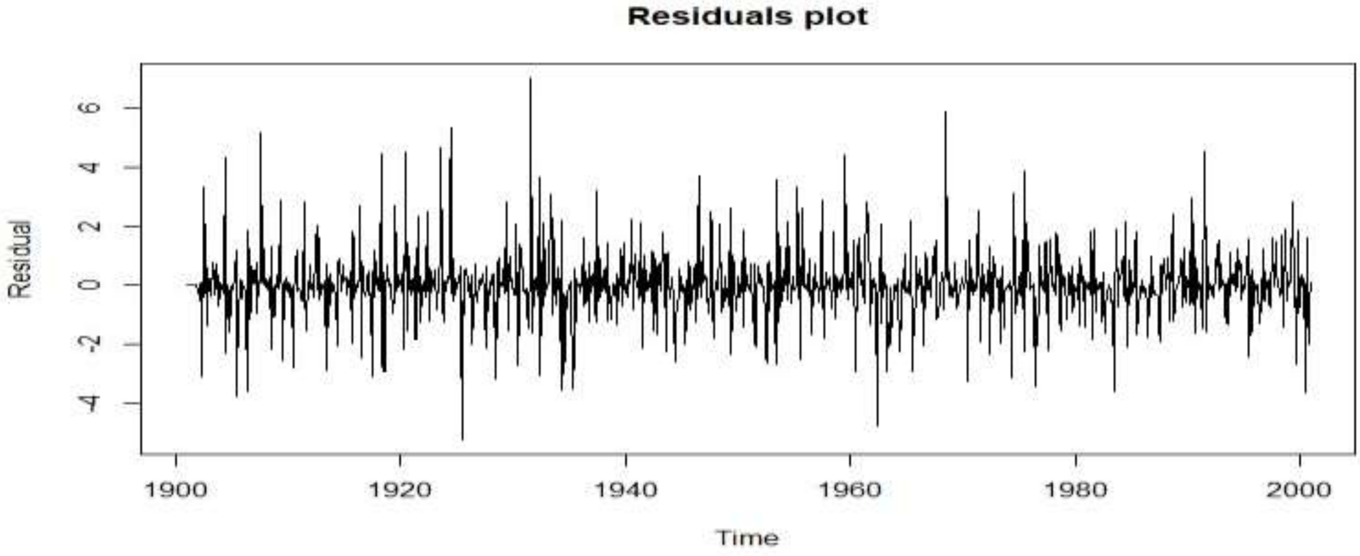
The model equation is

$$\boldsymbol{y_t = -0.0412 - 0.3255y_{t-1} - 0.5178y_{t-2} - 0.0377y_{t-3} - 0.6197y_{t-12} - 0.3121y_{t-24} + Z_t + 0.3398Z_{t-1} + 0.4996Z_{t-2}}$$

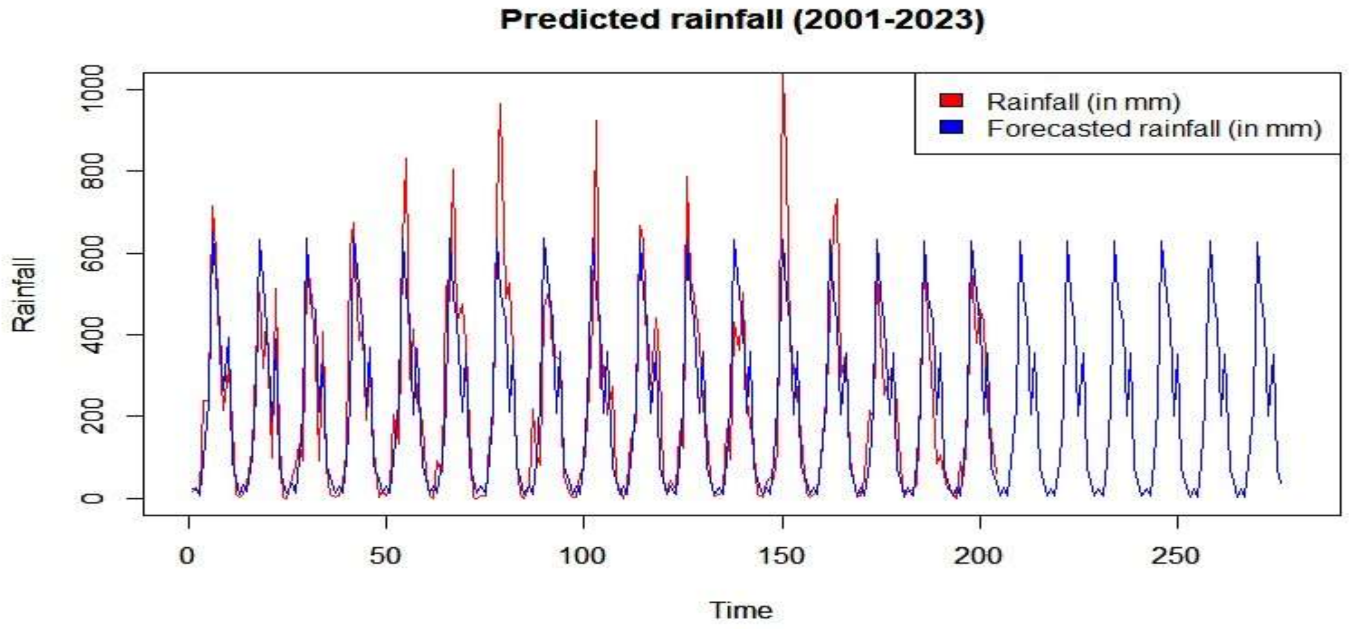
$$\boldsymbol{where\ \ Z_t\ is\ white\ noise}$$

Diagnostic check

We have calculated the standardized residuals and plotted them. As we can see from the graph, most of the values are lying in a band (evenly distributed above and below 0). So, we can conclude that the residuals have a constant 0 mean and constant variance.

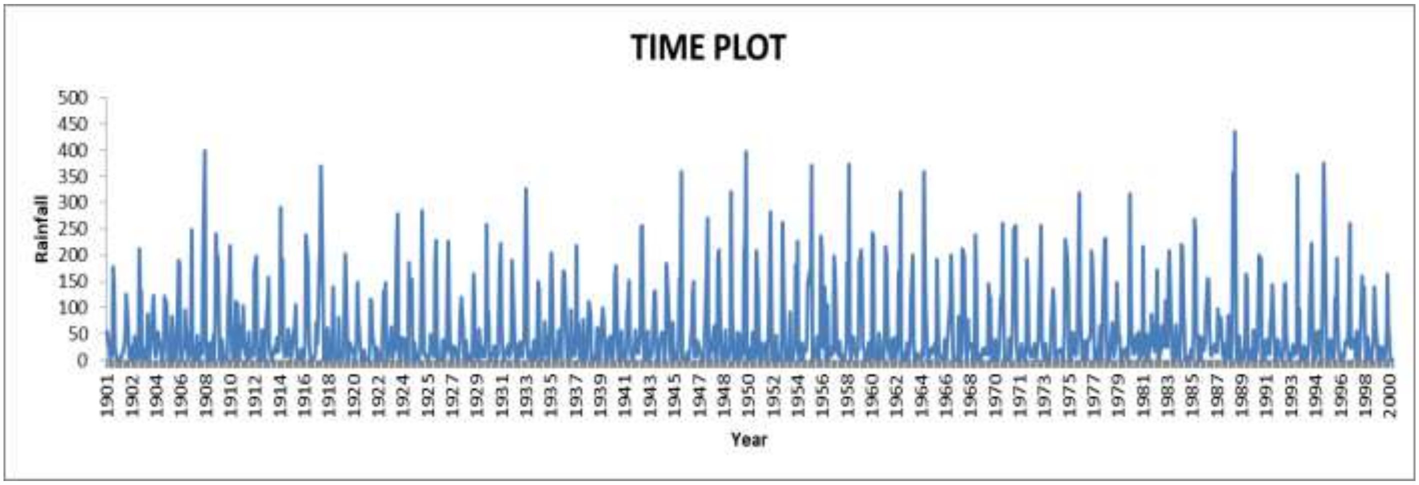


Forecasting:



2) PUNJAB

Stationarity

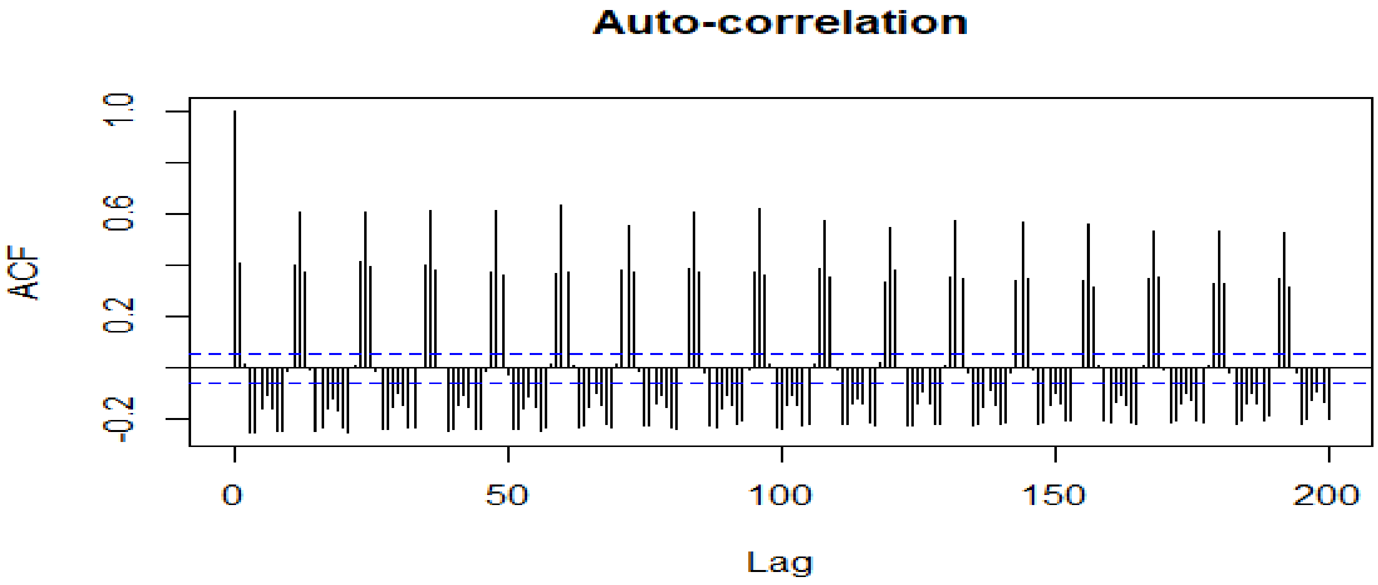


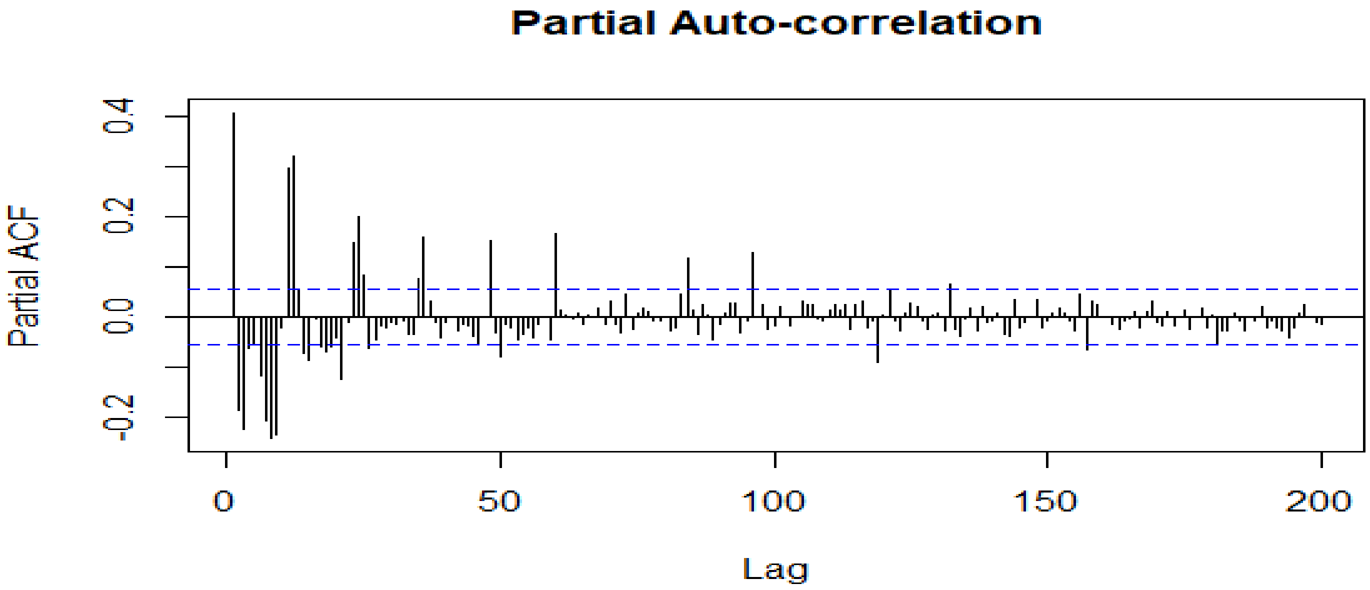
The time plot shows that data is mean and variance stationary and this was confirmed by Augmented Dickey-Fuller test.

H_0 : The time series is non-stationary
H_1 : The time series is stationary
Level of significance = 5%
Result: Dickey-Fuller = -12.706, Lag order = 10, p-value = 0.01

Since, p-value < 0.05, therefore, we reject the null hypothesis. Hence, we can conclude that the **time series is stationary**.

AC.F and PAC.F





From the ac.f plot, it is visible that there are significant spikes in the plot at lags with period=12. So, there is a seasonal component present which should be accounted while choosing a model for our data. Since, there are significant spikes in both ac.f and pac.f plots , so an ARIMA model with seasonal parameters i.e. **SARIMA model with no. of seasons = 12** will be the best suitable model for our data.

Model fitting

We fitted a SARIMA model on our data using auto.arima() function which gives us the model as:

ARIMA(0,0,0)(2,1,0)[12]

Coefficients:

	sar1	sar2
	-0.6797	-0.346
s.e.	0.0271	0.027

sigma^2 = 2670: log likelihood = -6374.51

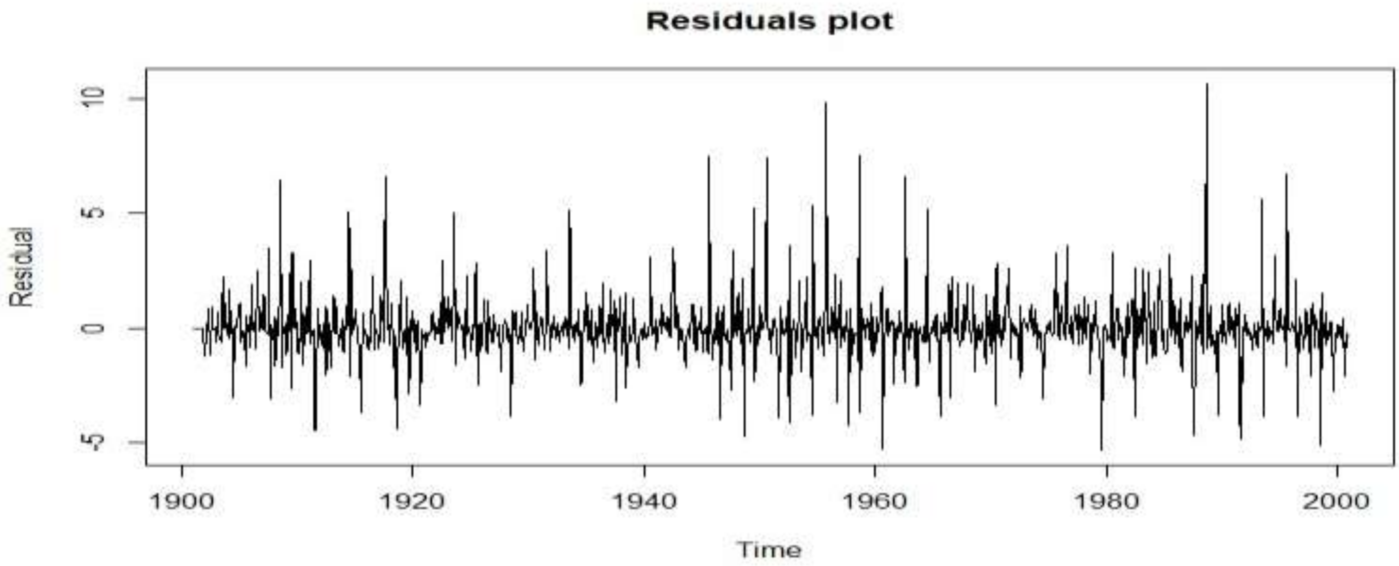
AIC=12755.02   AICc=12755.04   BIC=12770.26

The model equation is

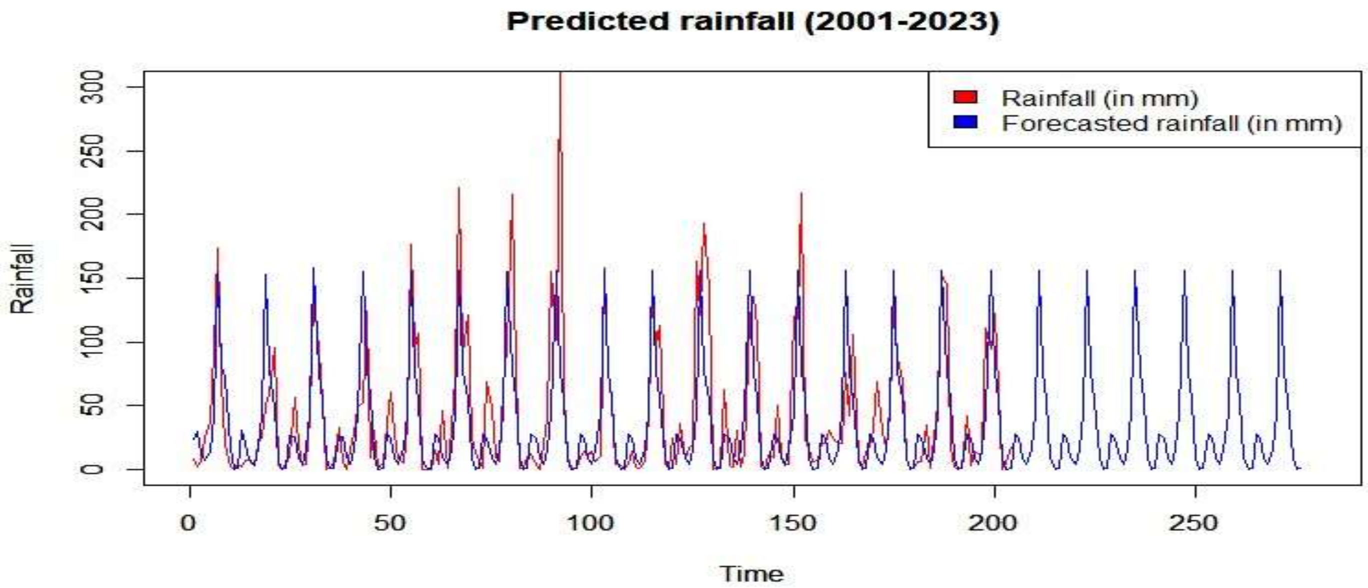
$$\boldsymbol{y_t = 0.6797y_{t-12} - 0.346y_{t-24}}$$

Diagnostic check

Most values of the standardized residuals are lying in a band (evenly distributed above and below 0). So, we can conclude that the residuals have a constant 0 mean and constant variance.



Forecasting



LIMITATIONS

1) The outliers present in the data require further analysis.
2) Cyclic component (assumed to be negligible) might be affecting the model fitting and forecast values.
3) Fine-tuning of parameters present in auto.arima() function in R might have resulted in a better fitting model.