TIME SERIES ANALYSIS

Seasonal & Non-Seasonal Time Series Analysis.

PROFESSOR: HADI SAFARI KATESARI

NAME: Shubham Neema

(CWID:20007579)

Table of Contents

NON-SEASONAL

DATASET -: Population of India (1950-2018)

Frequency: Yearly, Not Seasonally Adjusted

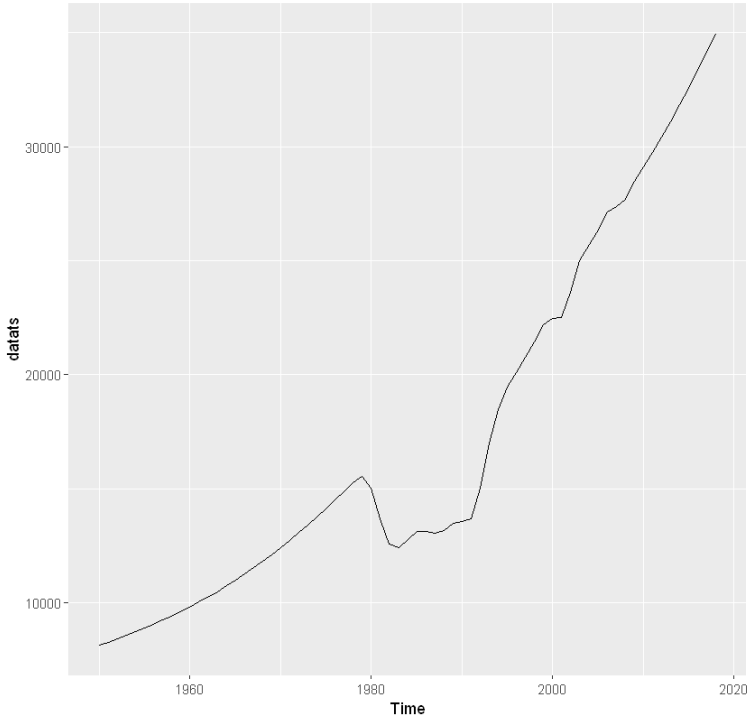
Introduction-:

In this research, we will examine a data set labeled "Indian Population from 1950 to 2018." India has world second largest population in the world , so it might be interesting to see per year population count and how and why did India become the second most populated country in the world

We want to build a model to explain the nature of the dataset's behavior using this data, and then test our suggested models to select the best one among them. We would be able to catch the significant autocorrelation at different lags in the data by evaluating ACF, which amounts to suggesting acceptable models. To define non-seasonal terms, we evaluate the early lags (1, 2, 3,...). Pins in the ACF (at low lags) show non-seasonal MA terms. Pins in the PACF indicate potential non-seasonal AR terms. The best model is one with a higher Log likelihood and a lower AIC criterion. We will also be able to test our models. This will be accomplished by taking into account the residuals' behavior.

Data Description and Analysis

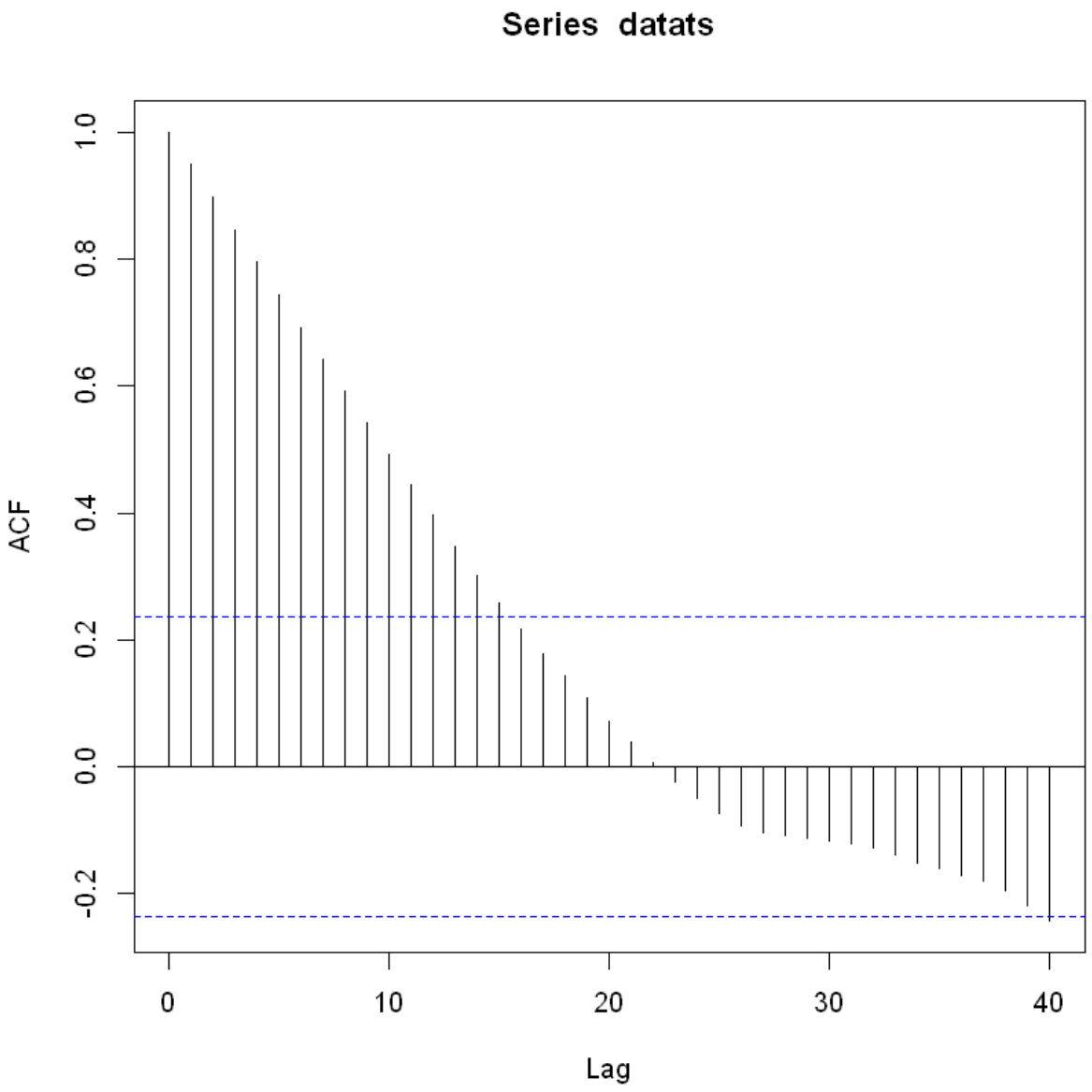
```
> #reading data from excel file

> Data_popl <- read_excel("populationIndia.xlsx")

> datats <- ts(Data_popl$Population,start = c(1950,1),frequency=1)

> autoplot(datats)
```



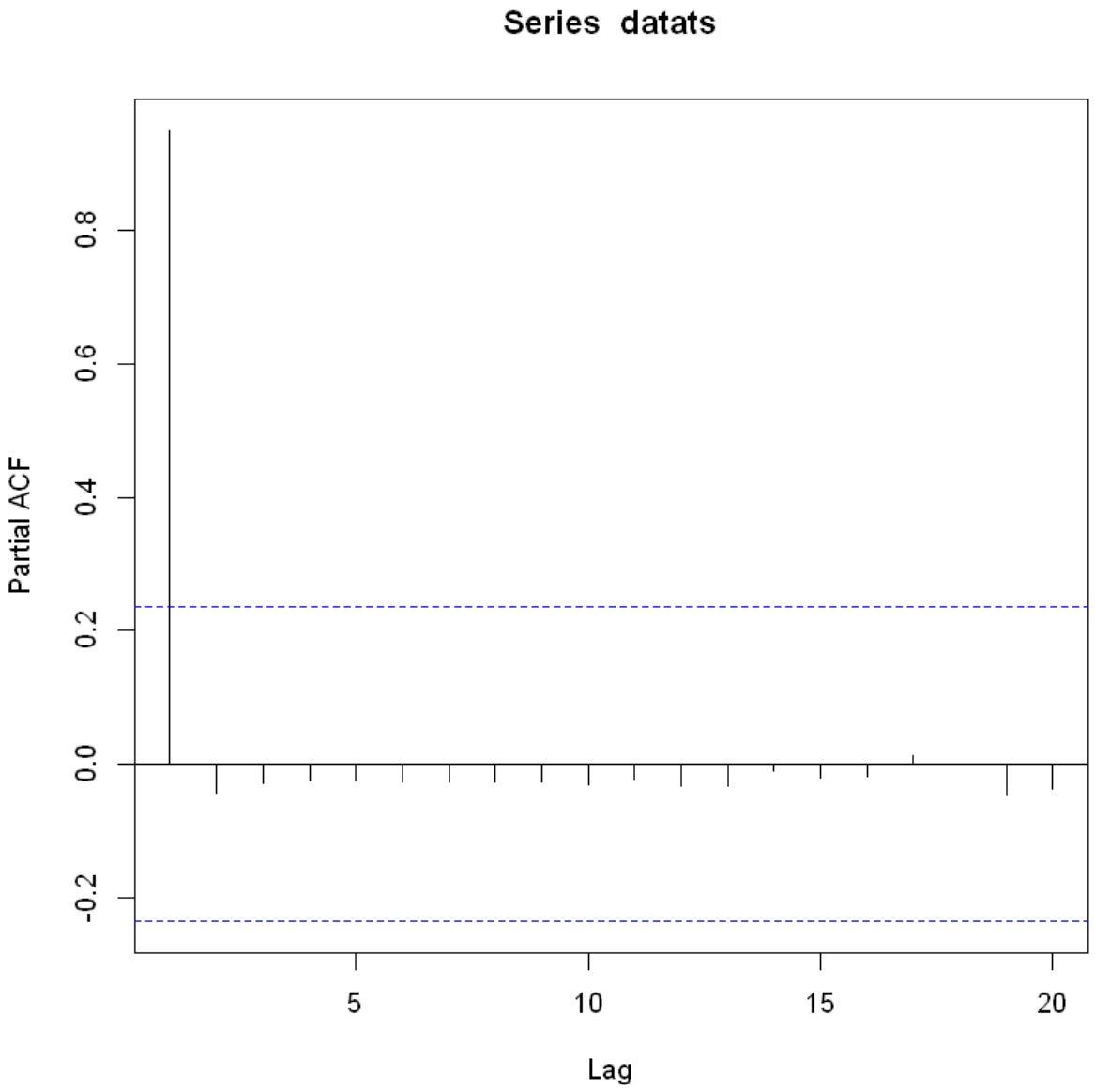
>

This graph depicts the annual population numbers from January 1950 to December 2018. There is a strong upward tendency from 1950 to 1980, but this is not the case from 1981 to 1991, when it fell for two years and additional growth was moderate until 1991. This tendency alone would necessitate the use of a nonstationary model. In the next plot, we can view better and with more information.

acf(datats,lag.max=40)



pacf <- pacf(datats,lag.max = 20,na.action = na.pass)



- **Stationary check**

adf.test(datats)

OUTPUT-:

Augmented Dickey-Fuller Test

data: log(datats)

Dickey-Fuller = -1.7527, Lag order = 4, p-value = 0.6757

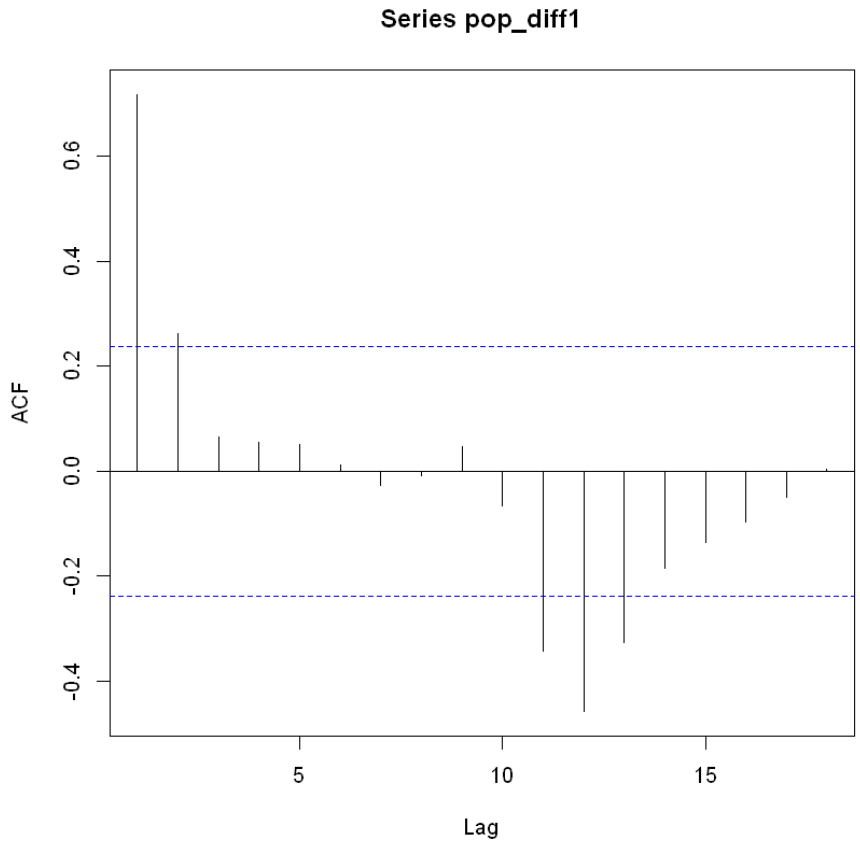
alternative hypothesis: stationary

After evaluating the ACF plot, we can see that there is still a good level of correlation among the data, and after performing the dicky fuller test, the p value was larger than

0.05, indicating that the data is nonstationary.

As a result, we will take a first-order difference and then perform the dicky fuller test again.

acf(pop_diff1)



The ACF plot of population data first differencing is shown above. There is a significant lag at 1 and relatively low autocorrelation in the data.

Again, the Dicky Fuller test is used to determine whether the data was stationary.

adf.test(pop_diff1)

Augmented Dickey-Fuller Test

data: pop_diff1
Dickey-Fuller = -2.5067, Lag order = 4, p-value = 0.3693
alternative hypothesis: stationary

Again p value is greater than 0.05 , so again we will apply differencing to make data stationary.

tem_diff2=diff(pop_diff1,differences=2)

adf.test(tem_diff2)

OUTPUT-:

Augmented Dickey-Fuller Test

data: tem_diff2

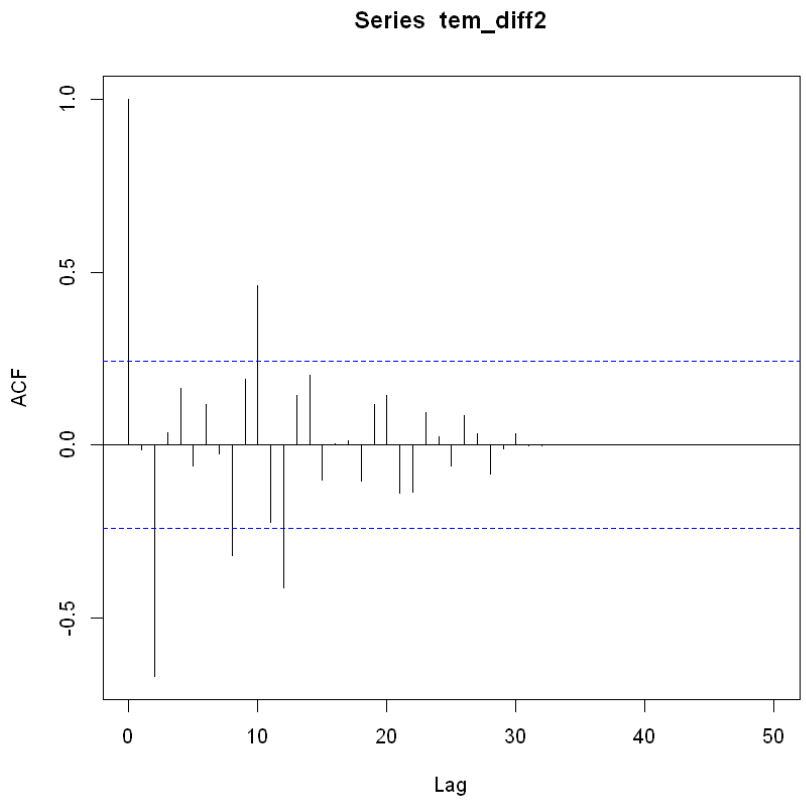
Dickey-Fuller = -5.7325, Lag order = 4, p-value = 0.01

alternative hypothesis: stationary

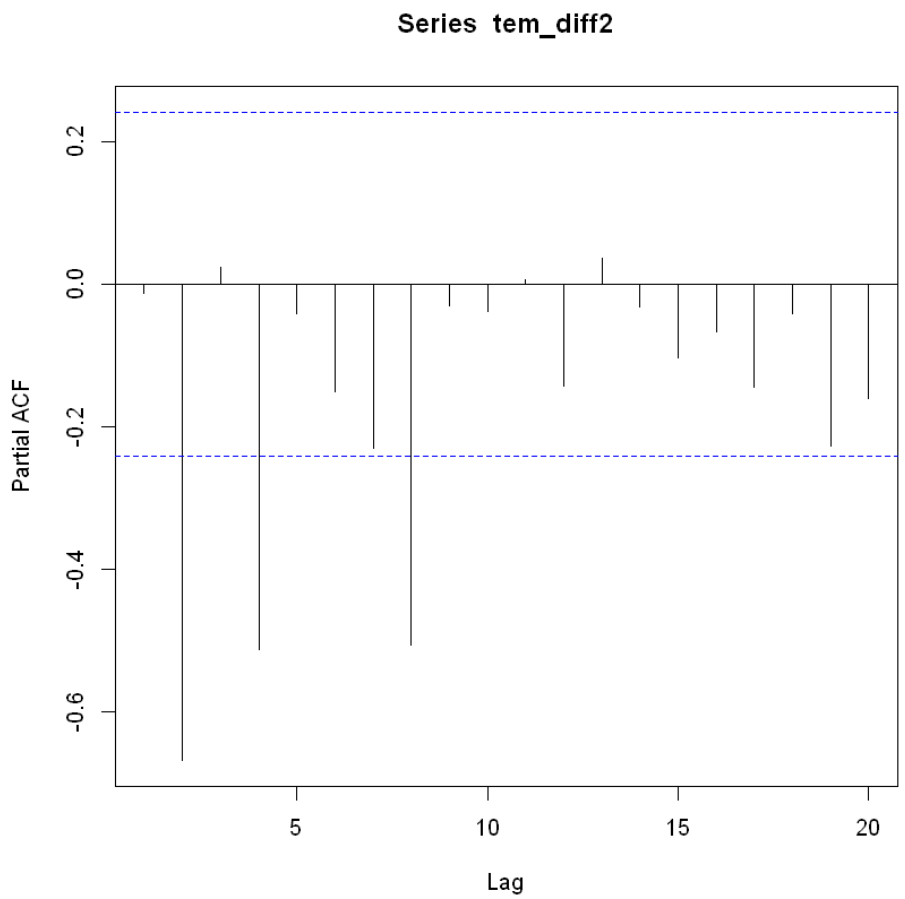
Here Data is finally stationary as p value is less than 0.05 ,hence we can reject the null hypothesis and take that the series is stationary.

Model Implementation and Evaluation-:

acf(tem_diff2,lag.max=50)



The plot of ACF (after two differences) shows that there is very little autocorrelation in the series after these two differences. We can notice a significant correlation at lags 2, 10, and 12 based on the ACF plot. Pins in the ACF (at three low lags) represent non-seasonal MA (3) terms for stationary data.



According to the plot of PACF (after two differencing), we can see strong correlation at lags 2, 4 and 8. Pins in the PACF (at 2 low lags) indicate non-seasonal AR (2) terms.

Here we have multiple values of p and q to check , it can be AR (2) (4) or (8) while MA (2) (10)(12) is also possible.

Amongst all the possible combinations ,ARMA(8,2,2) seems to be more promising as they have minimum AIC and BIC values and more log likelihood value.

fit <- Arima(tem_diff2, order=c(8,2,2))

Series: tem_diff2

ARIMA(8,2,2)

Coefficients:

	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8
	-0.1452	-1.1268	-0.2727	-0.9216	-0.3236	-0.6199	-0.2192	-0.4365
s.e.	0.1104	0.1056	0.1613	0.1561	0.1554	0.1540	0.1009	0.1024

	ma1	ma2

-1.9961 0.9997

s.e. 0.0881 0.0881

sigma^2 estimated as 64415: log likelihood=-451.28

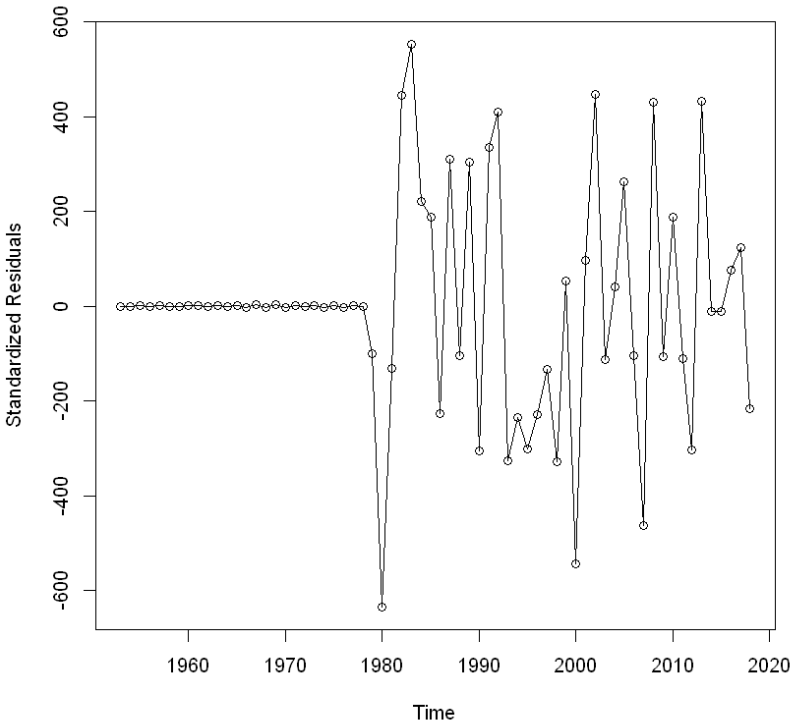
AIC=924.56 AICc=929.64 BIC=948.31

With having suggested model ARMA (8,2,2) we can proceed to parameter estimation of our model.

The coefficient estimates are all highly significant, and the next step is to check the model.

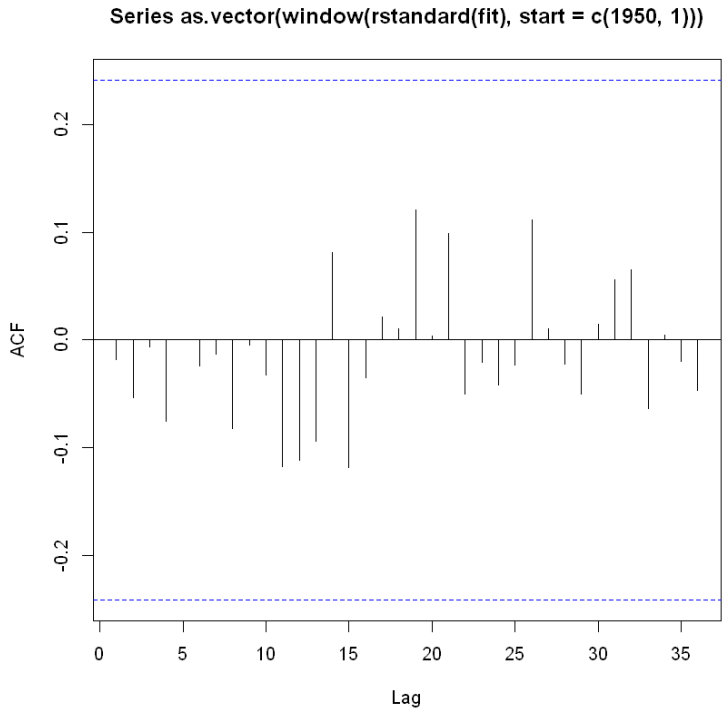
To check the estimated monthly ARIMA (8,2,2) model, we first look at the time series plot of the residuals.

```
plot(window(residuals(fit),start=c(1950,1)),ylab='Standardized Residuals',type='o')
```



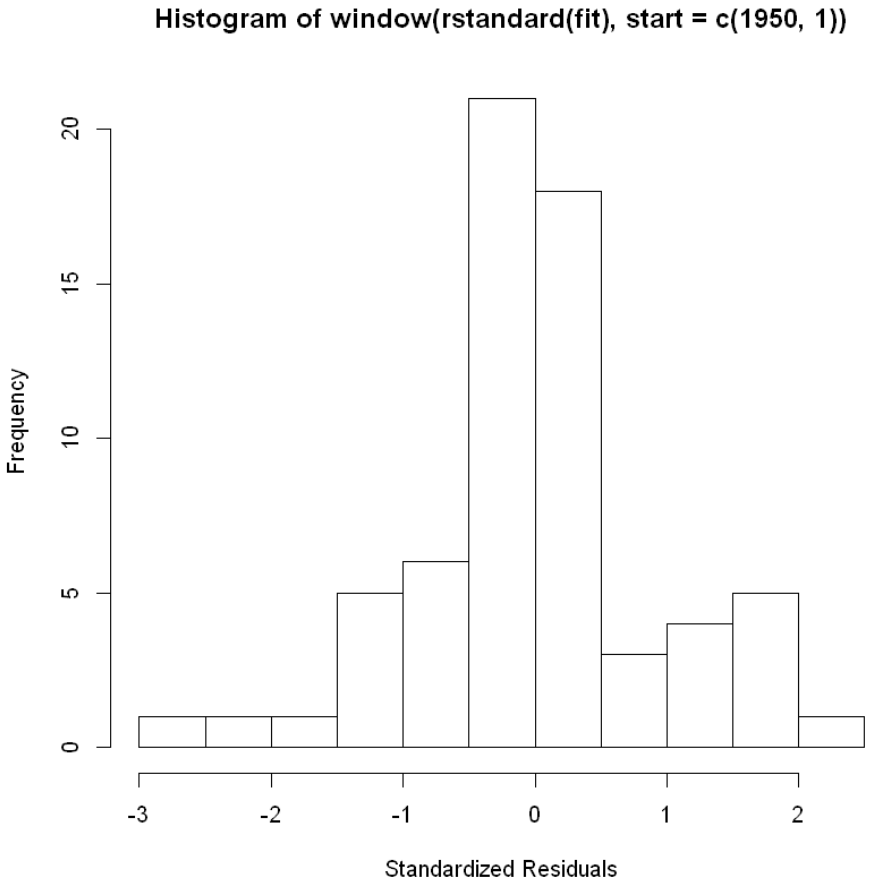
This plot gives us standardized residuals. The plot does not suggest any main disorder with the model .Also, we may need to consider the model further for outliers.

```
acf(as.vector(window(rstandard(fit),start=c(1950,1))),lag.max=36)
```



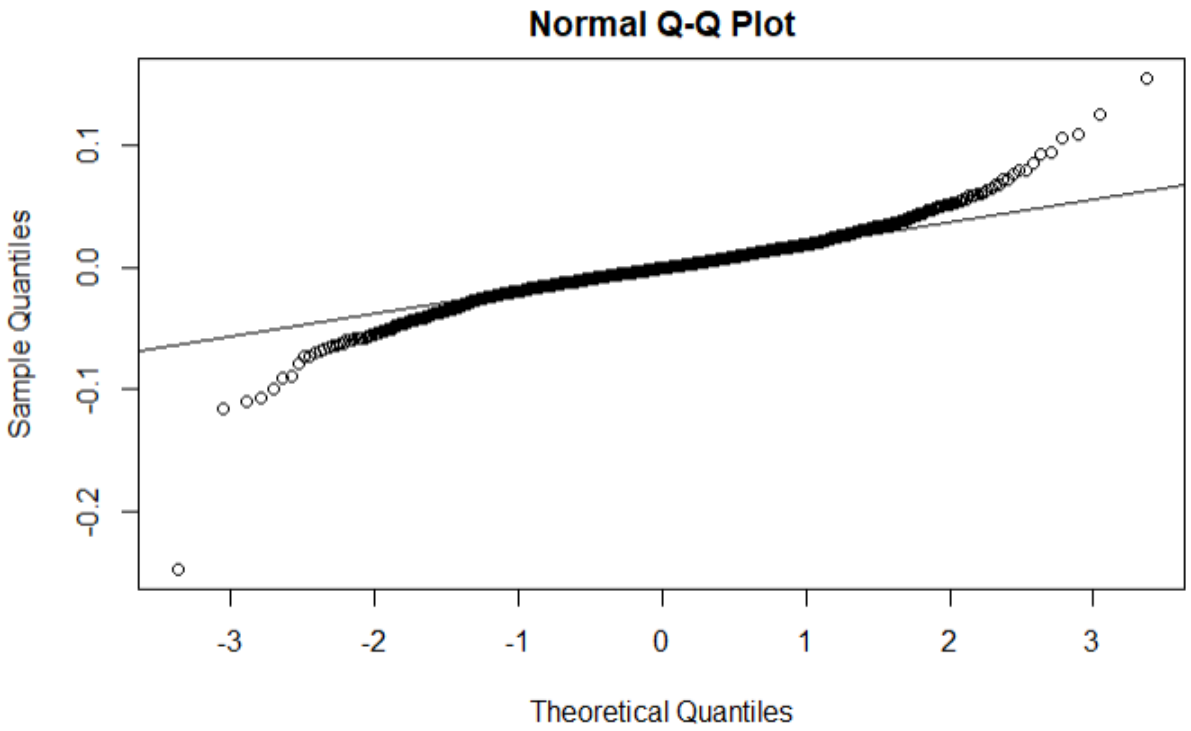
Here, we plot the sample ACF of the residuals. There isn't any “statistically significant” correlation here, which means there is no significant correlation present between the data.Also, we can consider normality of the error terms by the residuals.

```
hist(window(rstandard(fit),start=c(1950,1)),xlab='Standardized Residuals')
```



This plot demonstrates the histogram of the residuals. The shape of the plot is to some extent bell-shaped but definitely not exactly bell-shaped.

```
qqnorm(residuals(fit));
qqline(residuals(fit))
```



The quantile- quantile plot for the residuals from the ARIMA (8, 2, 2) Here the extreme values look suspect, while most of the values are closed to the center.

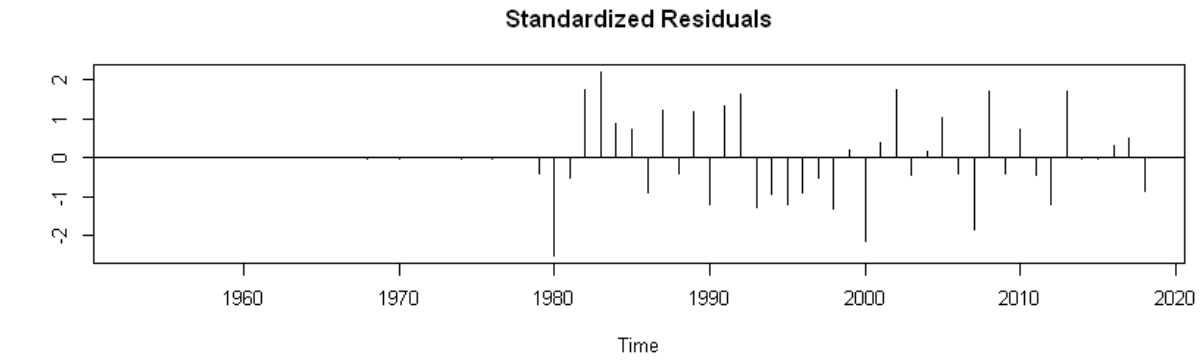
```
shapiro.test(residuals(fit))
```

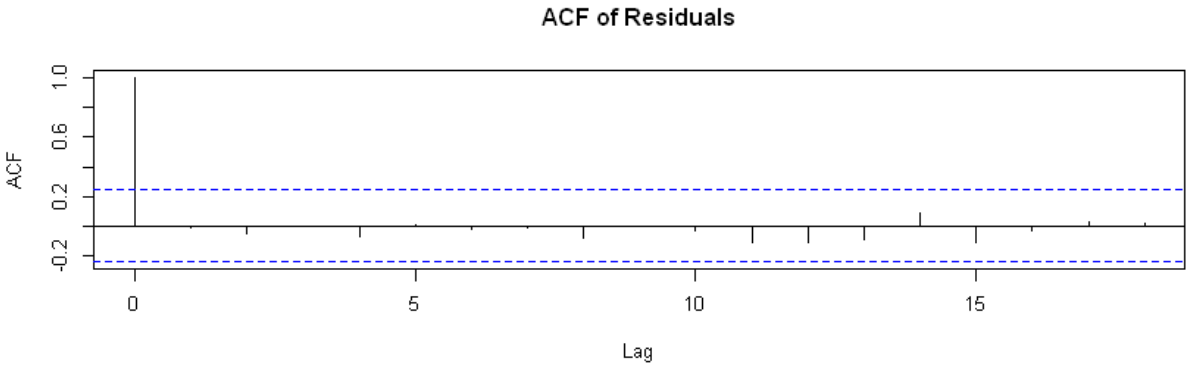
data: residuals(fit)

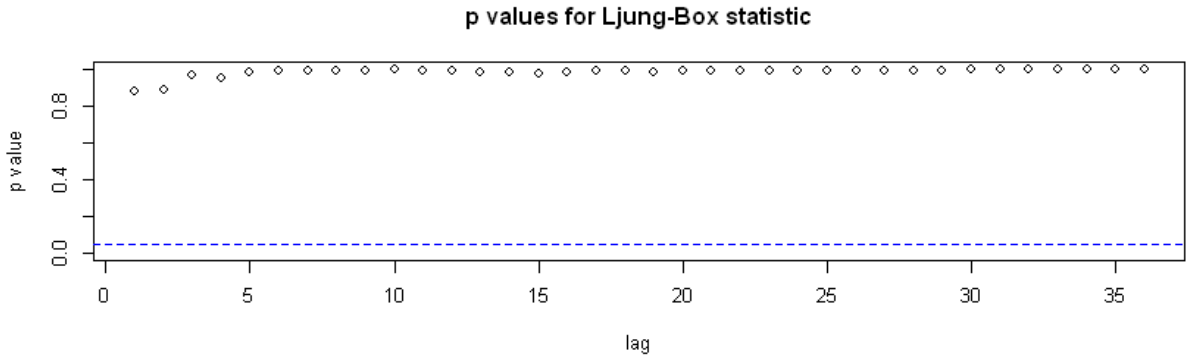
W = 0.97549, p-value = 0.2234

According to the result of output we can see, it has a test statistic of W = 0. 97549, leading to a p-value of 0. 2234. Thus normality is not rejected at any of the usual significance levels. Also, we can use Ljung-Box test for this model.

```
tsdiag(fit,gof=36,omit.initial=F)
```





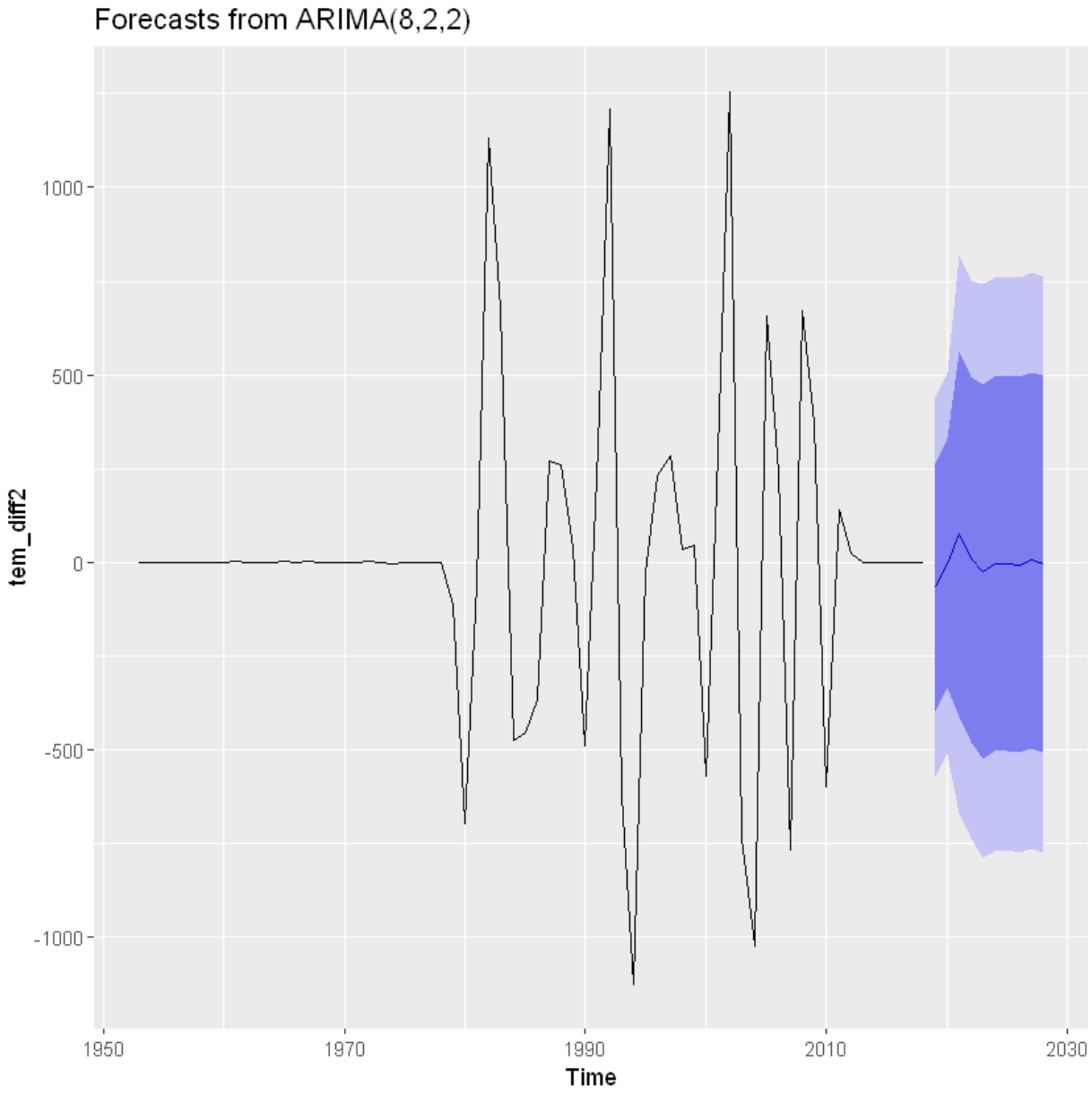


According to the Ljung-Box test for this model easily we can see a further indication that the model ARIMA (8, 2, 2) shows p value all above 0.05. It means we cannot reject null hypothesis and our residuals are random.

- **Forecast**-:

autoplot(forecast(fit))

.



This plot depicts the ARIMA forecasts and 95 percent forecast boundaries for a 12-year lead period (8,2,2). The projections closely reflect the stochastic periodicity in the data, and the forecast limits provide a strong sense of the forecasts' accuracy.

Conclusion

Overall, the forecast is reasonably accurate. The Ljung Box test reveals that the model residuals are non-autocorrelated, indicating that there is no heteroscedasticity problem and that the model is adequate; otherwise, we should investigate the GARCH model. The residuals of the model have a normal distribution and stationarity, indicating that the arima model fits the data well.

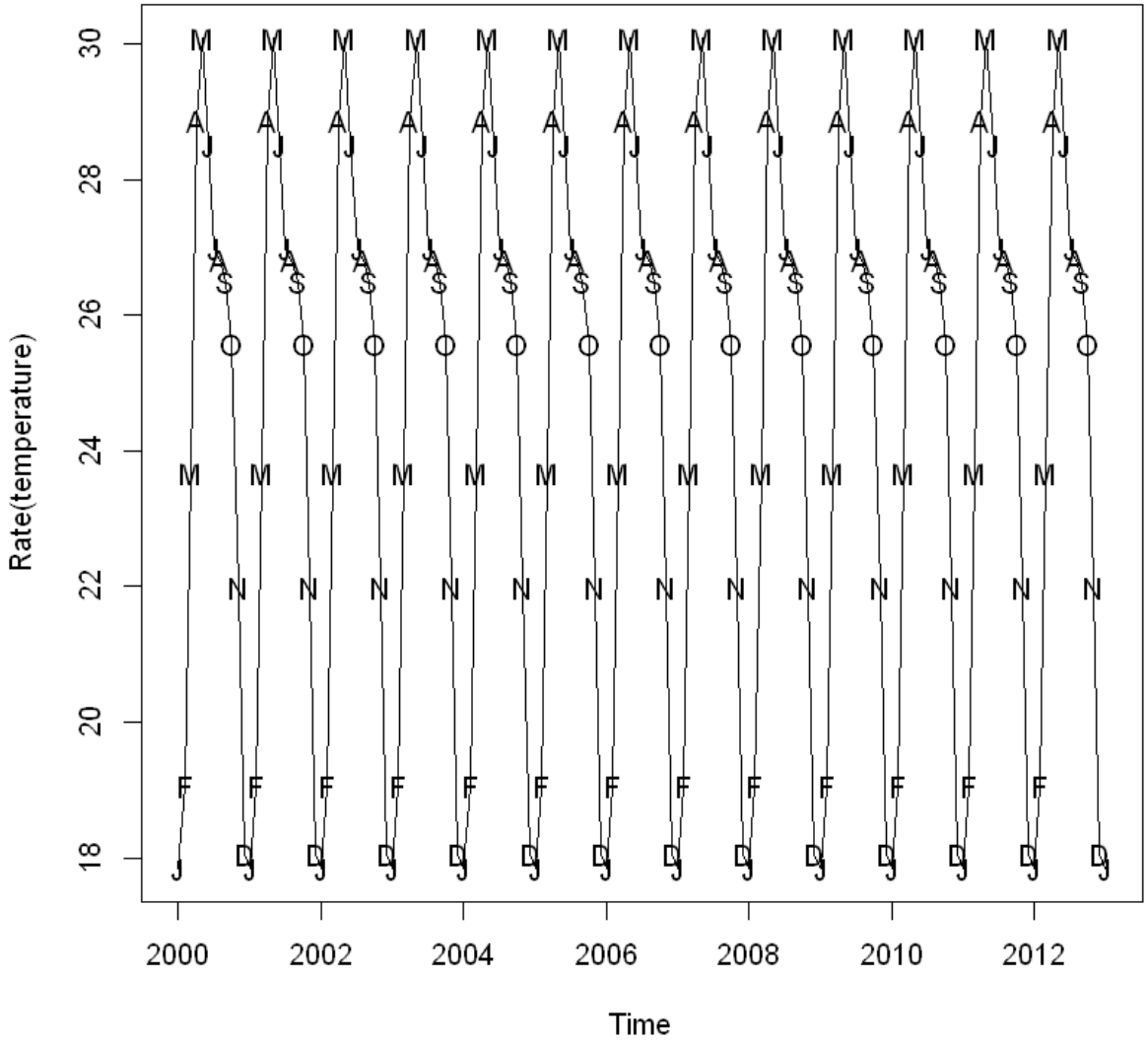
Seasonal Model

Introduction-

India’s land surface can be divided into six physiographic regions: Himalayan mountains in the north, Peninsular Deccan Plateau, the Indo-Gangetic Plains, Thar Desert in the west, Coastal Plain, and the Islands. All these regions have different climate profile and vulnerabilities.Given dataset has the average temperature of India from 2000 to 2012.We will try to understand the impact of global warming on India’s climate and also what will be the future looks like for india in terms of average temperature.

Data Description and Analysis

```
data<- read.csv("temperatureByCountry_India.csv")
datats=data$AverageTemperature
plot(window(datats,start=c(2000,1)),end=c(2013,1),ylab='Rate(temperature)')
Month=c('J','F','M','A','M','J','J','A','S','O','N','D')
points(window(datats,start=c(2000,1)),pch=Month)
```



Here we can clearly see seasonality ,and it make sense as this is average temperature and it should show seasonality. This plot demonstrates the monthly average temperature from Jan 2000 through Dec 2012.Every May , the temperature is increasing as it is summer at that time in India, similarly lowest in the month of december and january.This trend alone would lead us to specify a nonstationary model. We can see better and with the more details in the next plot

Stationary check-:

adf.test(tem_diff1)

Augmented Dickey-Fuller Test

data: tem_diff1

Dickey-Fuller = -9.9216, Lag order = 5, p-value = 0.01

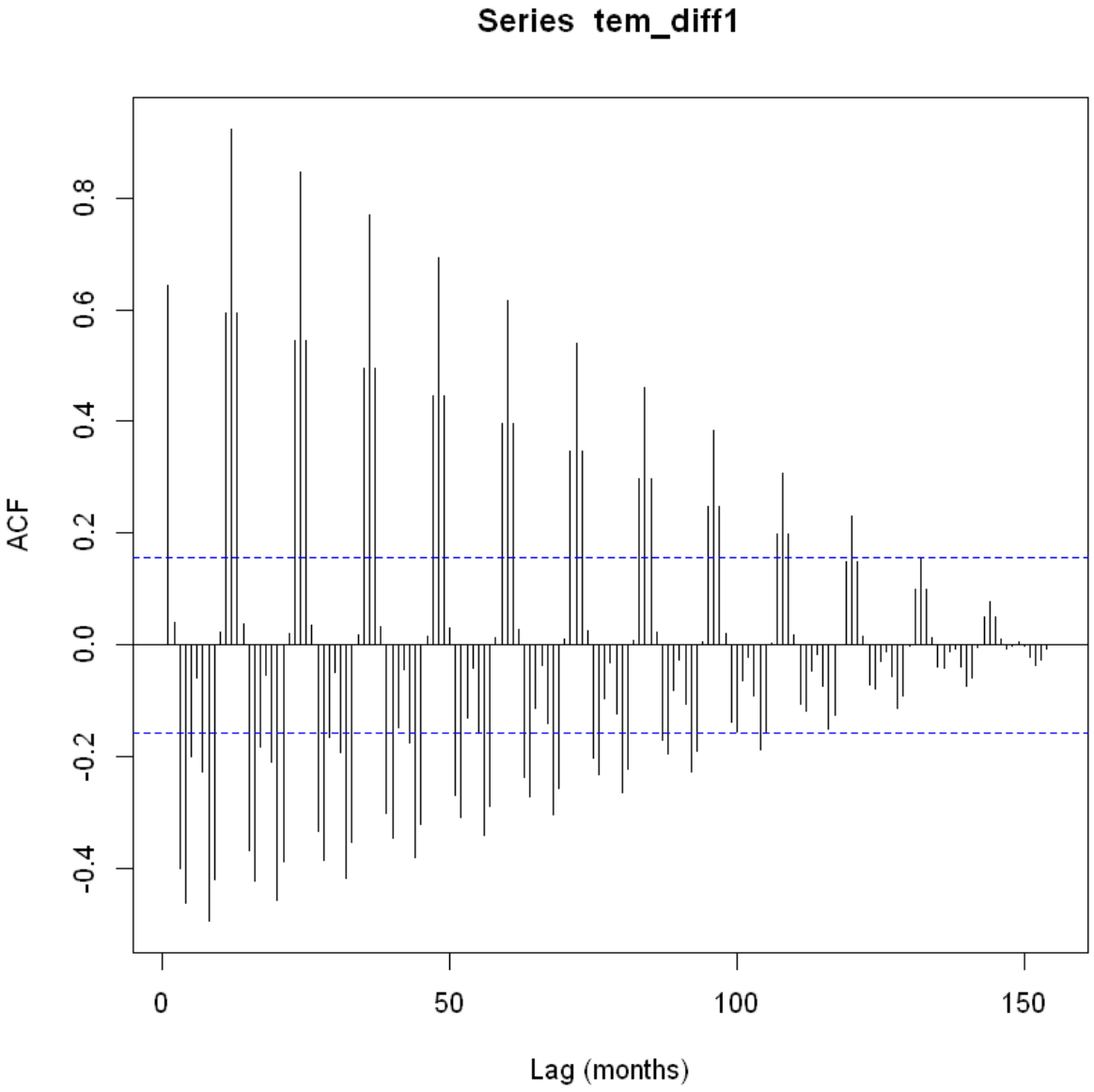
alternative hypothesis: stationary

P value is less than 0.05 , hence data is stationary.

- **Model Identification**

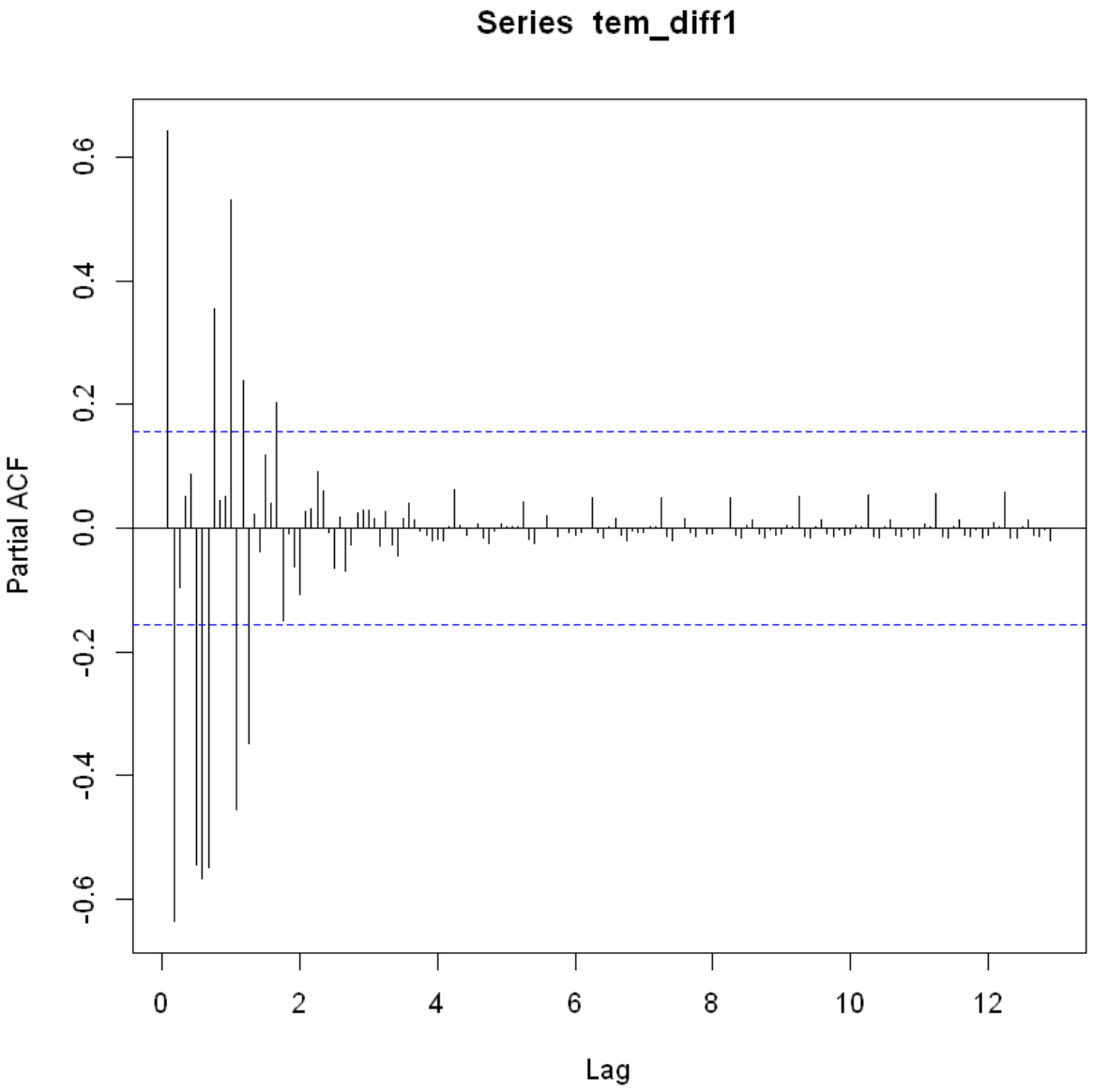
acf(tem_diff1, na.action = na.pass,lag.max=156)

We are taking acf of first order differencing data.



The plot demonstrates the sample ACF for our series. In this plot, the seasonal autocorrelation relationships are eye-catching, especially the strong correlation at lags 3,4,5,7,8,9,11 and etc.

pacf <- pacf(tem_diff1,lag.max = 156,na.action = na.pass)



According to the plot of PACF (after one differencing), we can see strong correlation at lags 2, ,6,7,8,9,10 and 12.

Various combinations are possible , hence best model will be decided based in AIC and Log likelihood value.

So, We will consider specifying , SARMA (7,1,8) × (0,1,2) 12

fit1=Arima(diff(datats), order=c(7,1,8),seasonal = list(order = c(0,1,2), period = 12),method="ML")

Series: diff(datats)

ARIMA(7,1,8)(0,1,2)[12]

Coefficients:

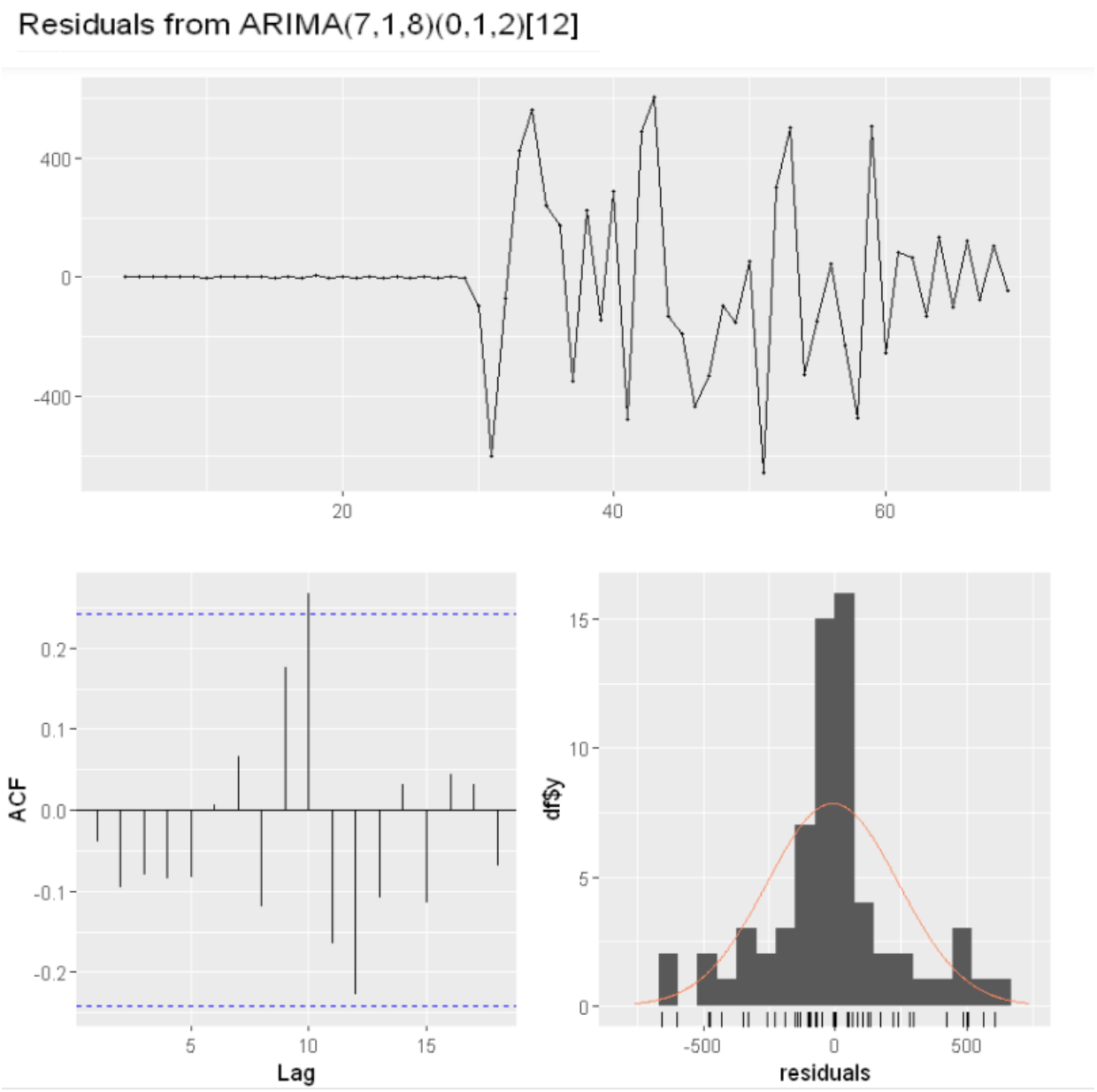
sigma^2 estimated as 6.787e-07: log likelihood=5011.07

AICc= 546.64   BIC=523.34

Out of all the possible combinations. ARIMA(7,1,8)(0,1,2)[12] has best AIC and log likelyhood value.Hence we will consider the same.

- **Model Implementation and Evaluation**

plot(window(residuals(fit1),start=c(2000,1)),ylab='Standardized Residuals',type='o')



Here only 1 significant lag at 12 , which can be false positive and hence can be ignored.

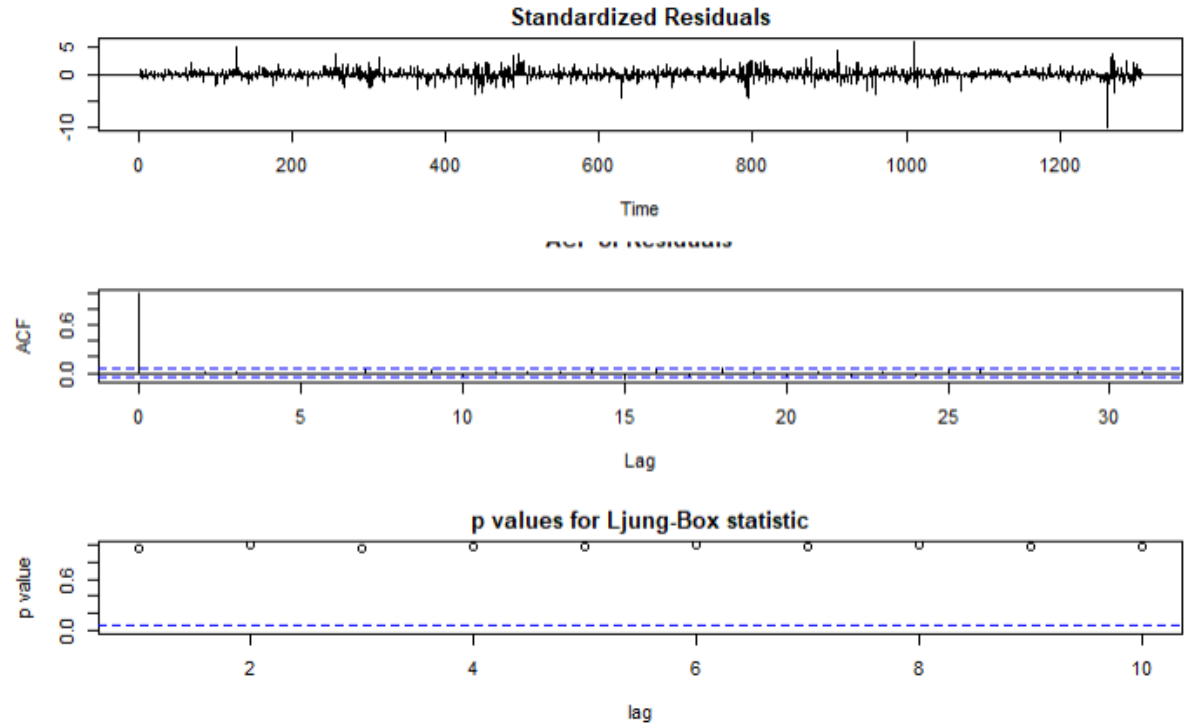
shapiro.test(residuals(fit1))

Shapiro-Wilk normality test

data: residuals(fit1)

W = 0.87598, p-value =0.024

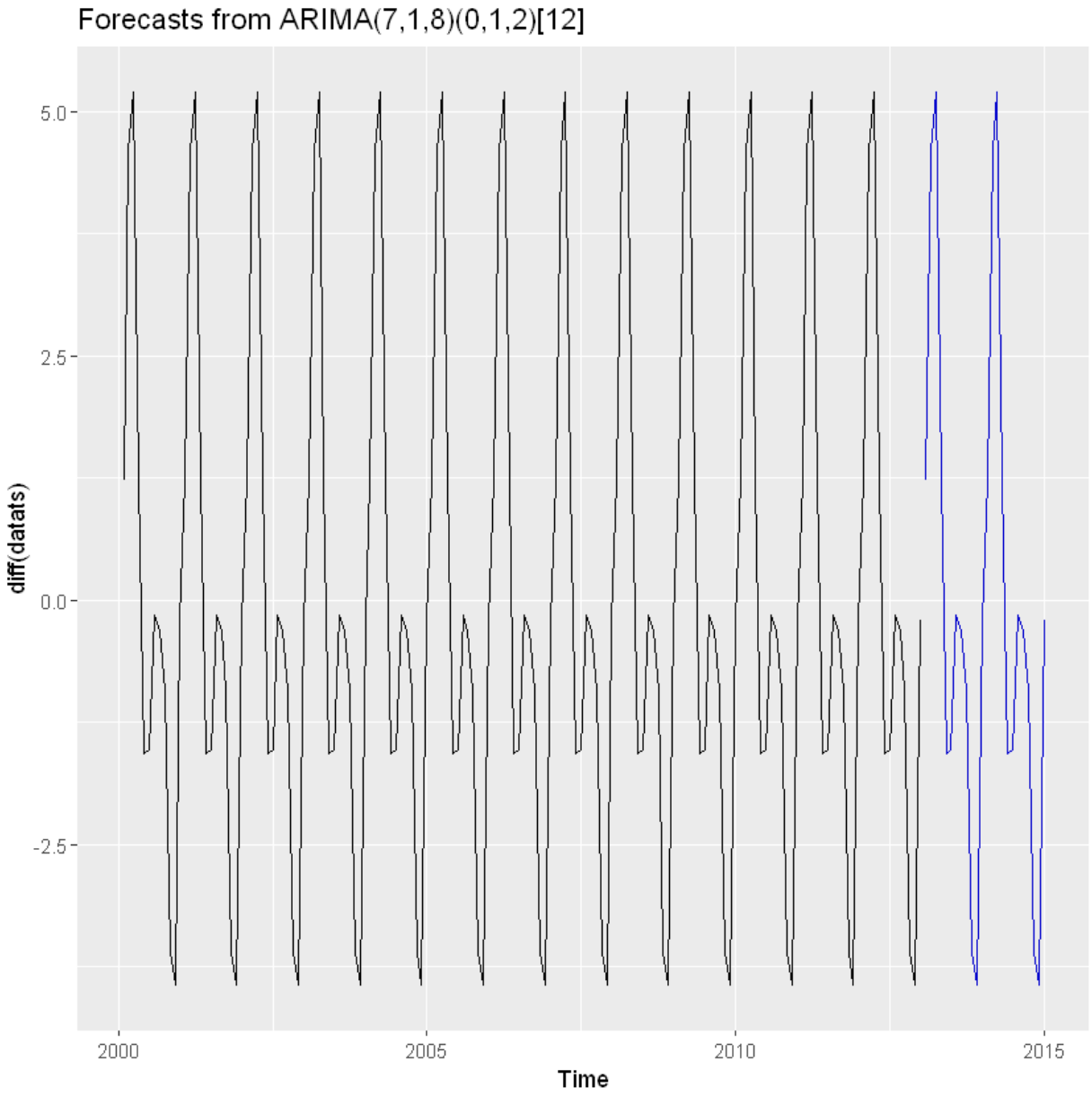
According to the result of output we can see, it has a test statistic of W = 0. 87598, leading to a p-value of 0.024. Thus normality is not rejected at any of the usual significance levels. Also, we can use Ljung-Box test for this model.



According to the Ljung-Box test for this model easily we can see a further indication that the model ARIMA(7,1,8)(0,1,2)[12] has grabbed the dependence in the time series.

Forecast

autoplot(forecast(fit))



This plot demonstrates the forecasts and 95% forecast limits for a lead time of 36 months for the SARMA (7,1,8) × (0,1,2) 12 model. The forecasts follow the stochastic periodicity in the data very well, and the forecast limits give a good feeling for the accuracy of the forecasts.

Conclusion-:

By looking at the forecast , significant change in temperature is not visible hence average temperature should not have much change in comparison to previous data.

Overall, the forecast is reasonably accurate. The Ljung Box test reveals that the modelresiduals are non-autocorrelated, indicating that there is no heteroscedasticity problem and that the model is adequate; otherwise, we should investigate the GARCH model. The residuals of the model have a normal distribution and stationarity, indicating that the sarima model fits the data well.