Time Series Analysis Project

Author: Prabhani Gunasekera

Contents

1.0 Time Series I

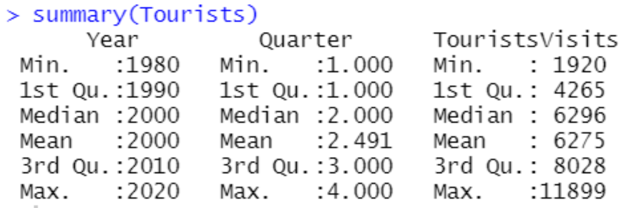
1.1 Data Set – Tourist Arrivals to UK (1980 to 2020)

For the first time series analysis, “UKTouristVisits” data set was obtained from Office for National Statistics website (*OS visits to UK:All visits Thousands-NSA* , 2020). It contains the number of tourists who visited UK from 1980 to 2020. The original data set contained yearly, quarterly, and monthly data. For this analysis, the quarterly dataset was chosen where the data is from first quarter of 1980 to first quarter of 2020.

1.2 Exploratory Data Analysis

Initially, as seen in figure 1.1 summary statistics of the data frame were checked.

The minimum of 1.92 million tourists in a quarter and a maximum of about 11.9 million tourists in a quarter. The average number of tourists visiting UK in a quarter between 1980 and 2020 had been approximately 6.3 million.



```
> summary(Tourists)
      Year         Quarter       TouristsVisits
 Min.   :1980   Min.   :1.000   Min.   : 1920
 1st Qu.:1990   1st Qu.:1.000   1st Qu.: 4265
 Median :2000   Median :2.000   Median : 6296
 Mean   :2000   Mean   :2.491   Mean   : 6275
 3rd Qu.:2010   3rd Qu.:3.000   3rd Qu.: 8028
 Max.   :2020   Max.   :4.000   Max.   :11899
```

Figure 1.1 – Summary Statistics of “UK Tourist Visits” Data

The time series plot in figure 1.2 below, shows a clear increasing trend in the data.

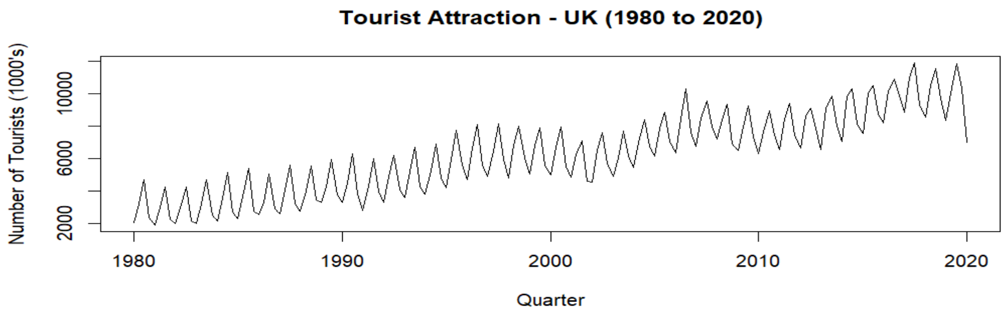


Figure 1.2 – Time Series Plot of “UK Tourist Visits” Data

To observe the points further a smaller window of time was taken and the below figure 1.3 was produced with Q1 to Q4 marked A-D.

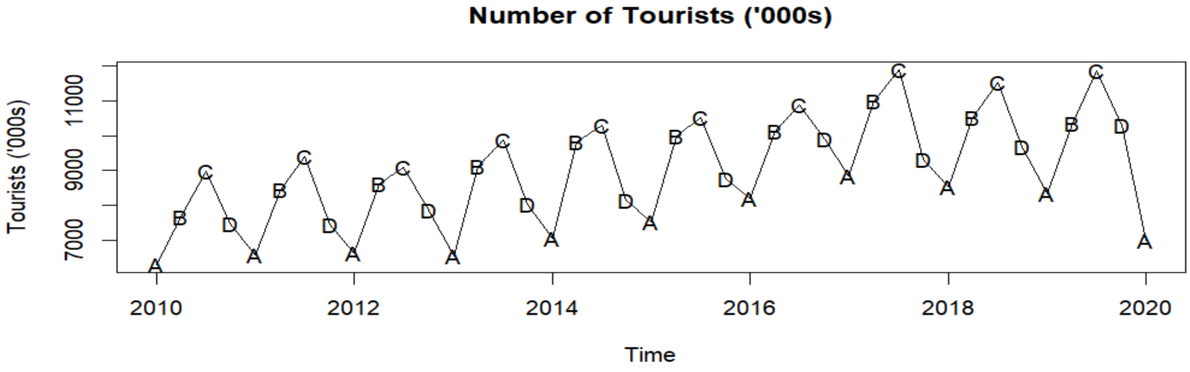


Figure 1.3 – Time Series Plot of “UK Tourist Visits” (2010-2020)

The time series oscillates within a constant band of variance, indicating an 'additive' model. With the position pattern of the letters A-D for the 4 quarters, it also clearly shows that a seasonal pattern exists within a year. This is further illustrated by the quarterly box plots in figure 1.4 below. Highest mean is seen in quarter 3, followed by quarter 2. This is possibly due to good weather conditions during those periods.

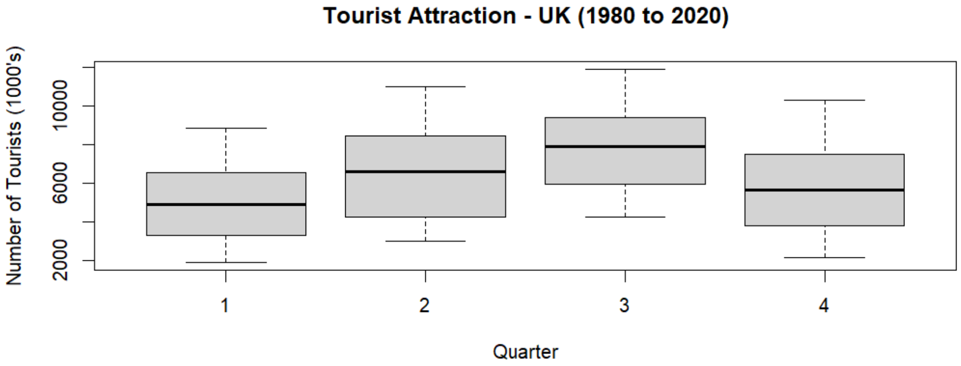


Figure 1.4 – Box Plot of Quarterly "UK Tourist Visits" (2010-2020)

The correlogram shown later in figure 1.7 of this report shows that the plot is decaying slowly with a wave pattern having peaks at every 4th lag, indicating the presence of trend and seasonality components in this time series data set. Therefore, differencing would be required before fitting the model to eliminate the trend and seasonality.

Figure 1.5 shows the time series decomposed under the additive method and the trend, seasonal component and random variation can be separately seen as follows.

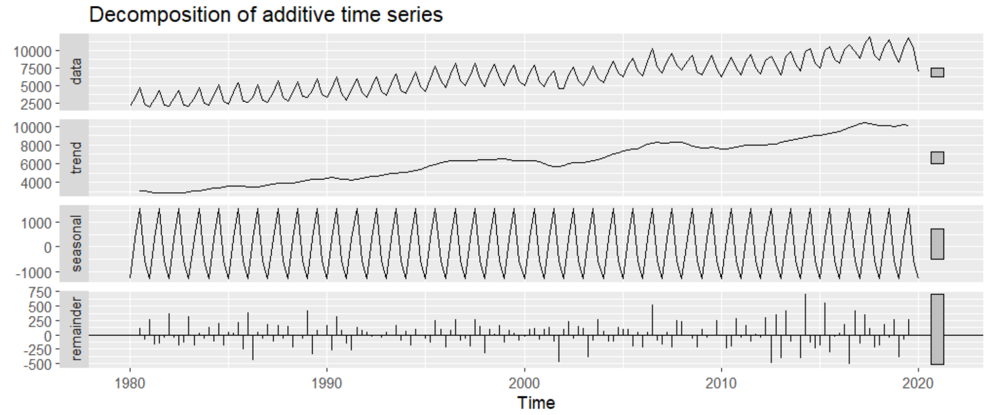


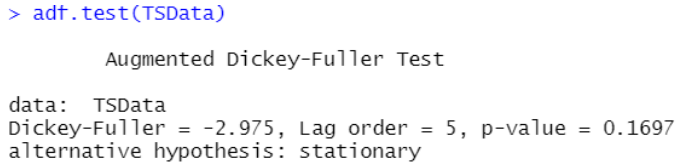
Figure 1.5 – Decomposition plot of the Tourist Visits time series data

1.3 Model Fitting

1.3.1 Check for Stationarity

The Augmented Dickey-Fuller (ADF) test was carried out to test the below hypothesis for stationarity. The test result is shown in figure 1.6.

H_o: The time series data is not stationary
H_1: The time series data is stationary



```
> adf.test(TSData)

        Augmented Dickey-Fuller Test

data:  TSData
Dickey-Fuller = -2.975, Lag order = 5, p-value = 0.1697
alternative hypothesis: stationary
```

Figure 1.6 – ADF Test for Stationarity for UK Tourist Visits Data

Here the p-value if greater than α = 0.05. Therefore, at 5% level of significance, we do not have enough evidence to reject the null hypothesis. Thus, we can conclude that the data is not stationary.

The stationarity can be further studied using the correlogram showing the autocorrelation function. Figure 1.7 shows the autocorrelation plot. As explained before in section 1.2, the combination of slow decreasing trend, and large autocorrelations at the quarterly seasonal lags indicate the presence of both trend and seasonal components. Hence, the data is not stationary.

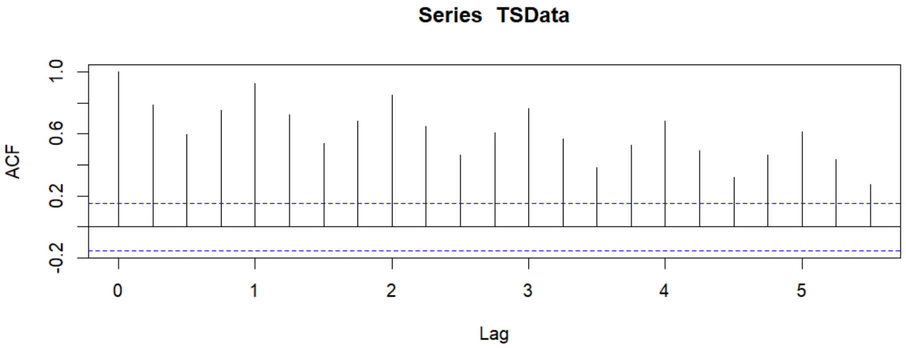


Figure 1.7 – Correlogram of UK Tourist Visits Data (1980-2020)

1.3.2 Removing Trend

To remove the trend component the difference of the original time series data set was taken. Figure 1.8 shows the plot of the difference time series. The trend appeared to be eliminated. However, the seasonal component was still present.

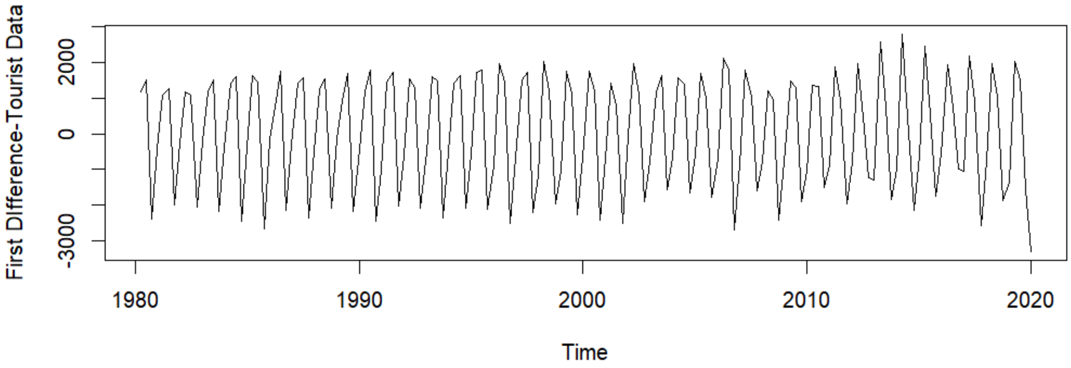


Figure 1.8 – First Difference of UK Tourist Visits Data (1980-2020)

The ACF of the differenced time series was also taken to further examine this. As per Figure 1.9 the trend appears to have been eliminated. However, the peaks at seasonal lags indicate that the seasonal component is present.

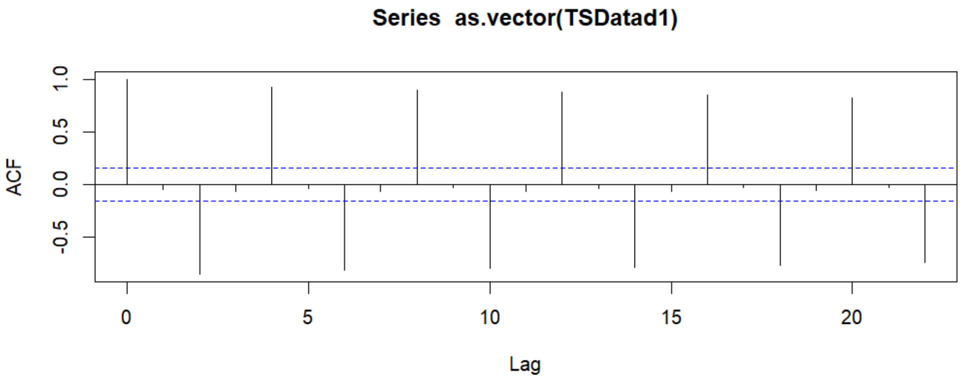


Figure 1.9 – ACF of First Difference of UK Tourist Visits Data (1980-2020)

1.3.3 Removing Seasonality

The seasonal difference of the data was taken to remove the seasonal component. Figure 1.8 shows the plot of the seasonal difference time series. The seasonal values seem to be more random here compared to figures 1.3 and 1.8.

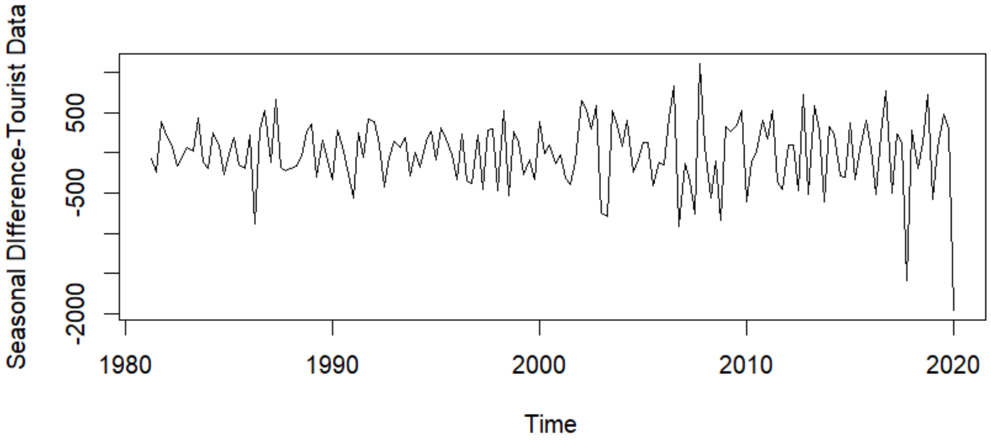


Figure 1.10 – Seasonal Difference of UK Tourist Visits Data (1980-2020)

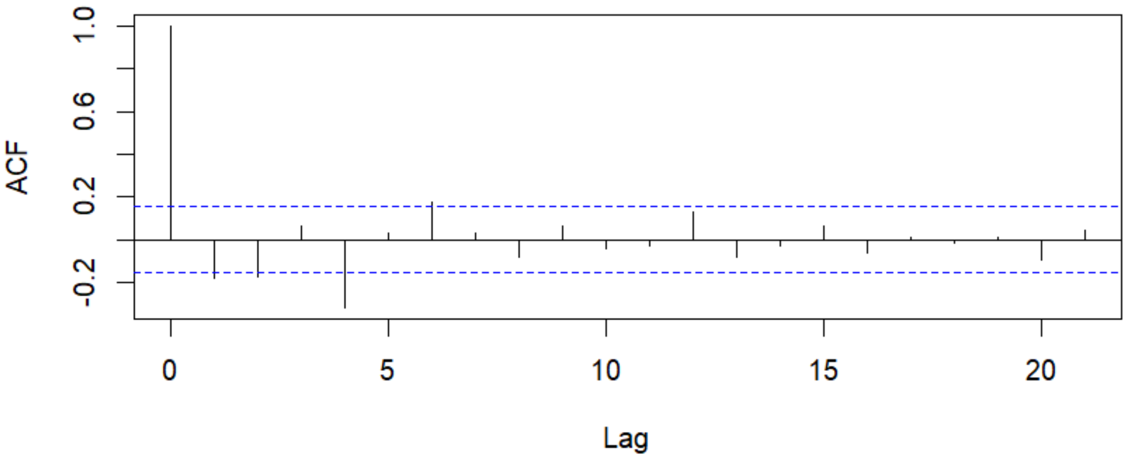


Figure 1.11 – ACF of Seasonal Difference of UK Tourist Visits Data (1980-2020)

As per the ACF of second difference in figure 1.11, most of the ACF values are within the interval of confidence, except for a few. Hence, the seasonality has been removed.

This shows the presence of a seasonal moving average because of the significant peak at 4th lag. There is no clearly significant peak indicating a normal moving average. Thus, it could be either MA(0) or MA(1) and the Q value would be 1.

Figure 1.12 shows the partial autocorrelation function. Significant peaks are again observed at 1st, 2nd and 4th lags. This suggests that an AR(1) or AR(2) model with P= 1.

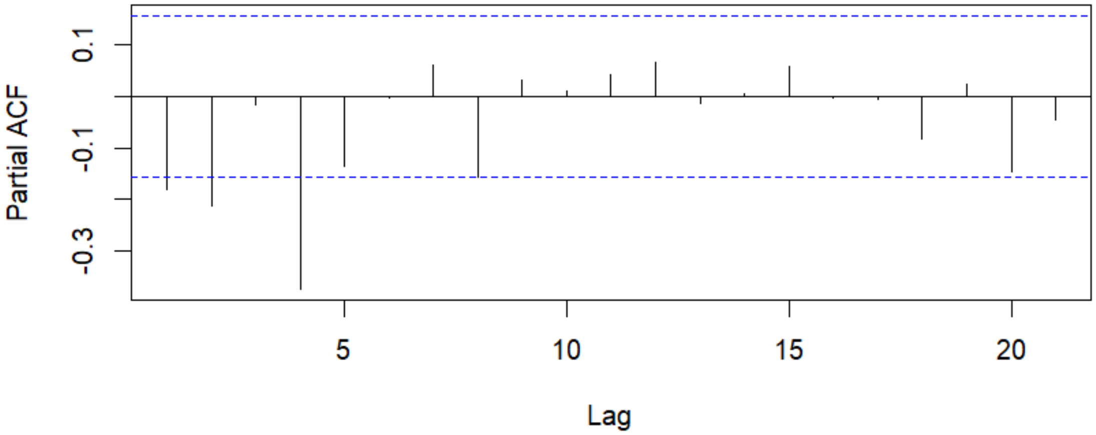
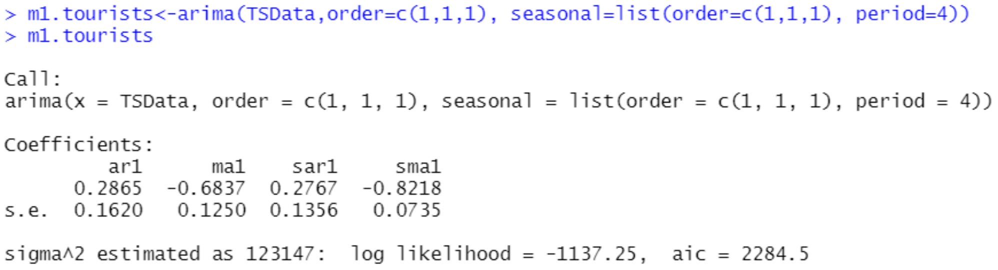


Figure 1.12 – Partial Autocorrelation Function of UK Tourist Visits Data (1980-2020)

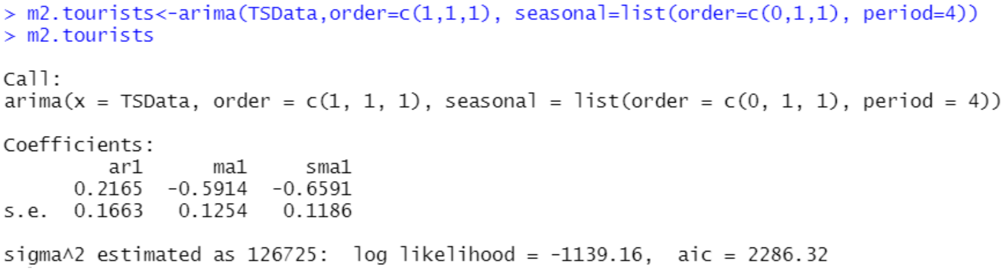
1.3.4 Parameter Estimates for the Model

Model 1 with ARIMA(1,1,1)(1,1,1) and model 2 with ARIMA(1,1,1)(0,1,1) were checked.



```
> m1.tourists<-arima(TSData,order=c(1,1,1), seasonal=list(order=c(1,1,1), period=4))
> m1.tourists

Call:
arima(x = TSData, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1), period = 4))

Coefficients:
         ar1      ma1    sar1     sma1
      0.2865  -0.6837  0.2767  -0.8218
s.e.  0.1620   0.1250  0.1356   0.0735

sigma^2 estimated as 123147:  log likelihood = -1137.25,  aic = 2284.5
```

Figure 1.13 – Model 1



```
> m2.tourists<-arima(TSData,order=c(1,1,1), seasonal=list(order=c(0,1,1), period=4))
> m2.tourists

Call:
arima(x = TSData, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 4))

Coefficients:
         ar1      ma1     sma1
      0.2165  -0.5914  -0.6591
s.e.  0.1663   0.1254   0.1186

sigma^2 estimated as 126725:  log likelihood = -1139.16,  aic = 2286.32
```

Figure 1.14 – Model 2

By using aic criterion, when compared the aic values in figure 1.13 and 1.4 above, model 1 is better than model 2. Therefore, the ARIMA(1,1,1)x(1,1,1)4 model is chosen as the best model. However, the residuals of this model must be checked for assumptions. The residuals must be a white noise.

1.3.5 Residuals Check

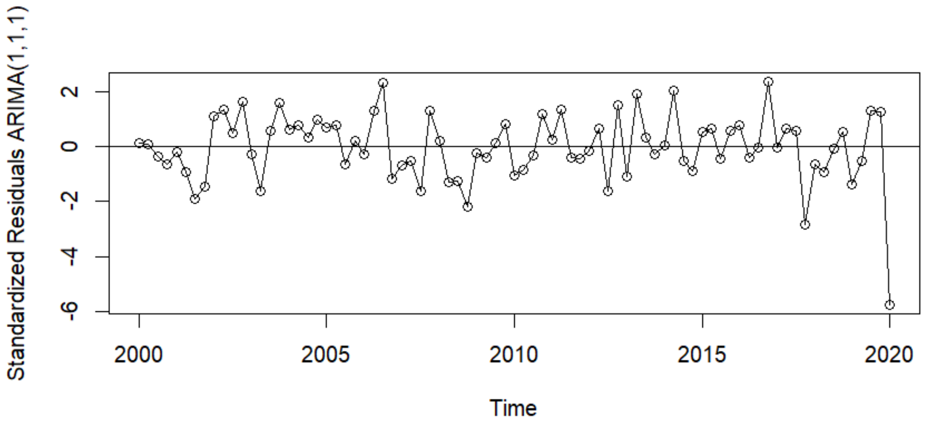


Figure 1.15 – Residuals plot of ARIMA(1,1,1)x(1,1,1)4

As per figure 1.15 the residuals look fairly alright as they are very close to zero except for one outlier in 2020 Q1 possibly due to an external factor.

The ACF of this model in figure 1.16 shows that ACF of all the residuals are within the interval of confidence. Hence, it appears that residuals are uncorrelated.

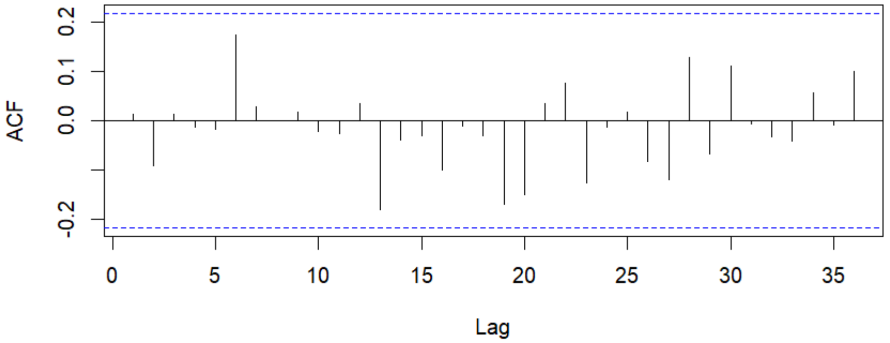
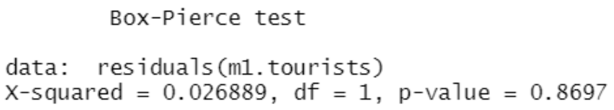


Figure 1.16 – ACF of Residuals of ARIMA(1,1,1)x(1,1,1)4

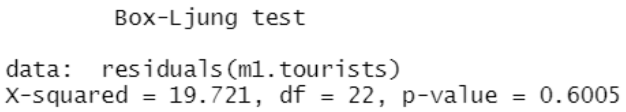
Box-Pierce test and Box-Ljung test were performed to check this further. The below hypothesis was tested.

H_o: The error terms are uncorrelated
H_1: The error terms are correlated



```
        Box-Pierce test

data:  residuals(m1.tourists)
X-squared = 0.026889, df = 1, p-value = 0.8697
```

Figure 1.17- Box-Pierce Test

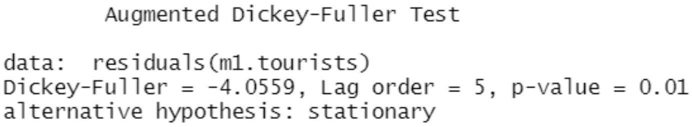


```
        Box-Ljung test

data:  residuals(m1.tourists)
X-squared = 19.721, df = 22, p-value = 0.6005
```

Figure 1.18- Box-LjungTest

As per both the figures 1.17 and 1.1.8 above the p-values are greater than α=0.05, indicating that the null hypothesis cannot be rejected. Hence, at 5% level of significance we have enough evidence to say that the error terms are uncorrelated. Thus, the residuals do not depend on time.

Thereafter, the ADF test was also performed to check the stationarity of the residuals. Figure 1.19 shows the output. Here, the p-value is less than α=0.05. Therefore, at 5% level of significance we reject the null hypothesis that the residuals are not stationary. This shows that the residuals are stationary.

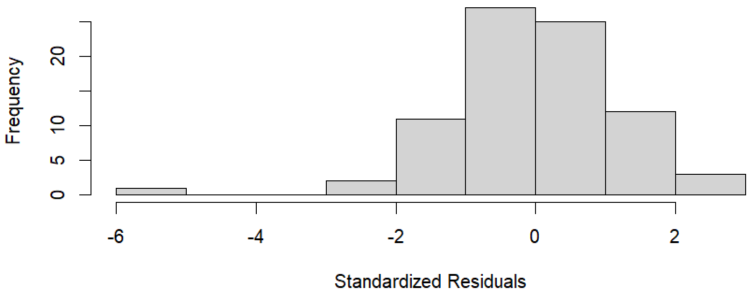


```
                Augmented Dickey-Fuller Test

data:  residuals(m1.tourists)
Dickey-Fuller = -4.0559, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 1.19- ADF Test for Residuals

Therefore, we can conclude that the residuals follow a white noise.

The residuals must also follow a normal distribution. Thus, the following plots were checked.



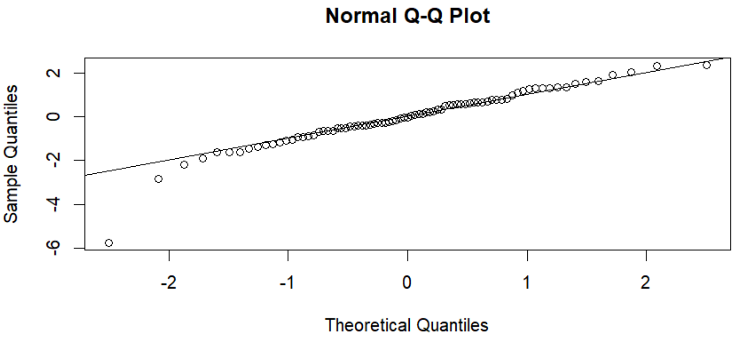
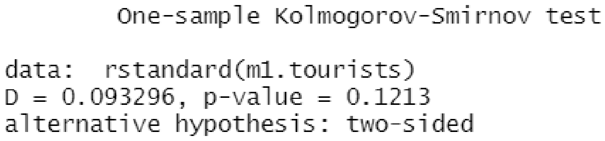


Figure 1.20 - Histogram of Residuals

Figure 1.21 – Q-Q Plot

The histogram of residuals (Figure 1.20) and Q-Q plot (Figure 1.21) both show that the residuals follow a normal distribution.



```
        One-sample Kolmogorov-Smirnov test

data:  rstandard(m1.tourists)
D = 0.093296, p-value = 0.1213
alternative hypothesis: two-sided
```

Figure 1.22 Kolmogorov-Smirnov Test

Kolmogorov-Smirnov test was done to further check the normality, and its output in Figure 1.22 shows that p-value is greater than α = 0.05, indicating the null hypothesis that residuals follow a normal distribution cannot be rejected at 5% level of significance.

Thus, the chosen model ARIMA(1,1,1)x(1,1,1)4 can be concluded to be the best model.

1.4 Forecasting

The forecasts plot for the next 4 quarters is shown in figure 1.22. It appears to be in line with the pattern in the previous quarters. Hence, by first looks the forecasts appear to be correct.

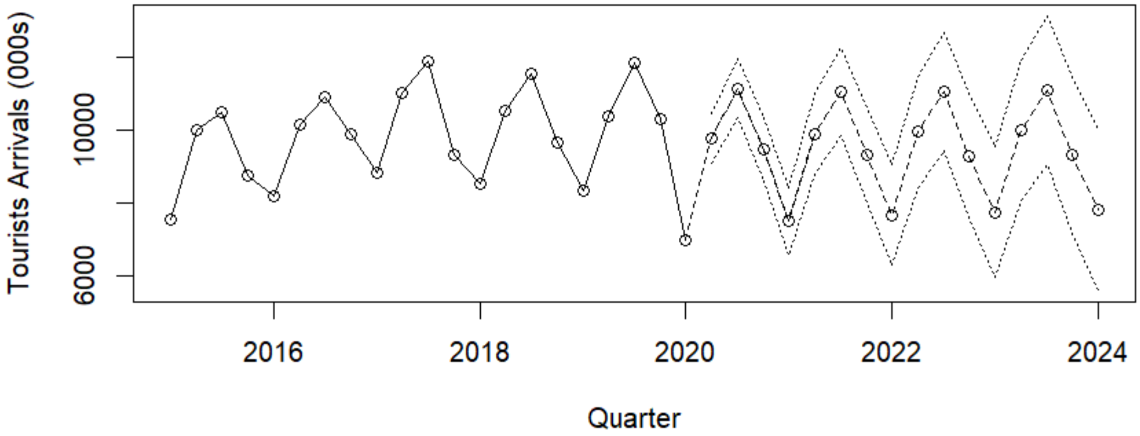
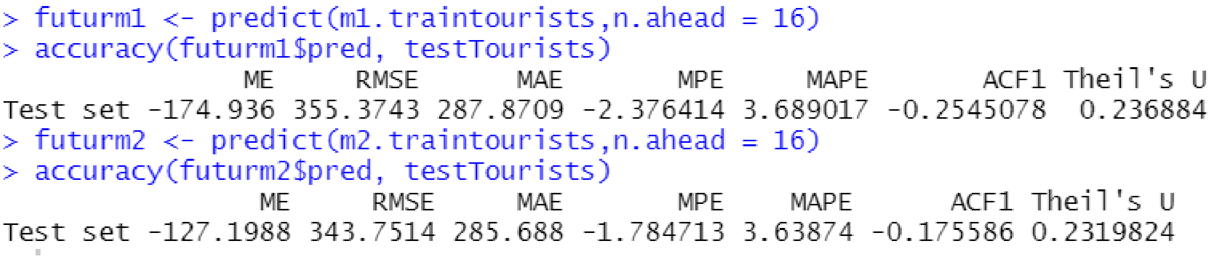


Figure 1.22 Forecasts for next 4 quarters

The data were then split into train and test sets.

Forecast accuracy was checked for both the models ARIMA(1,1,1)x(1,1,1)4 and ARIMA(1,1,1)x(0,1,1)4 by comparing the predictions with the test data set. The below figure 1.23 gives the accuracy output.



```
> futurm1 <- predict(m1.traintourists,n.ahead = 16)
> accuracy(futurm1$pred, testTourists)
               ME     RMSE      MAE       MPE     MAPE       ACF1 Theil's U
Test set -174.936 355.3743 287.8709 -2.376414 3.689017 -0.2545078  0.236884
> futurm2 <- predict(m2.traintourists,n.ahead = 16)
> accuracy(futurm2$pred, testTourists)
                ME     RMSE     MAE       MPE    MAPE      ACF1 Theil's U
Test set -127.1988 343.7514 285.688 -1.784713 3.63874 -0.175586 0.2319824
```

Figure 1.23 Forecasts Accuracy

As per the above output, the forecasts of the ARIMA(1,1,1)x(0,1,1)4 model seem to have lesser mean squared error and root mean squared error than ARIMA(1,1,1)x(1,1,1)4. Hence, it would be better to use the model ARIMA(1,1,1)x(0,1,1)4 for forecasting.

2.0 Time Series II

2.1 Data Set – Sunspots

"Sunspots" data set was obtained from Kaggle.com (Kaggle, no date). It contains the monthly mean sunspots form January 1749 to January 2021.

2.2 Exploratory Data Analysis

The minimum of number of sunspots in a month was zero while the maximum was 398.2. The average of mean monthly sunspots between 1749 and January 2021 had been approximately 82.

The time series plot in figure 2.1 below, shows a no clear trend in the data.

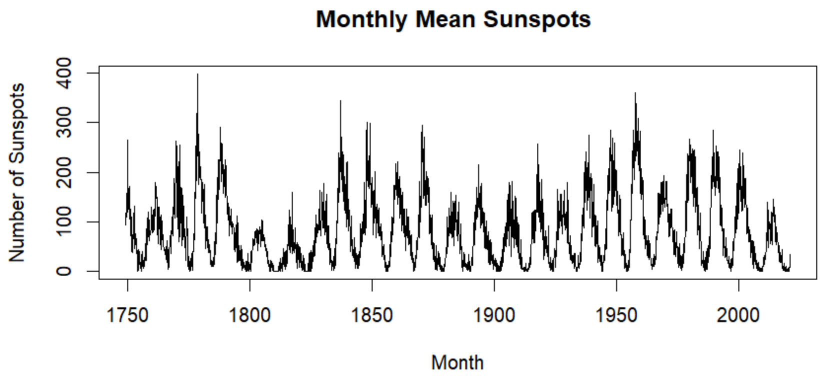


Figure 2.1 – Time Series Plot of Mean Monthly Sunspots Data

The time series oscillates within fairly a constant band of variance, indicating an 'additive' model. It shows that a seasonal pattern may exist. However, this must be further explored.

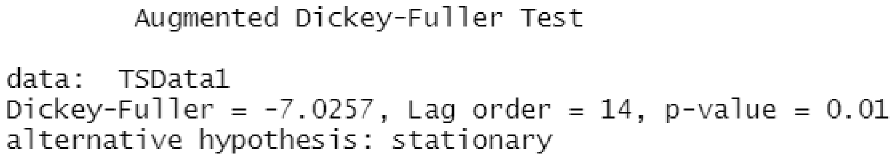
The correlogram shown later in figure 2.3 of this report shows that the plot is decaying fast with no pattern, indicating the possibility of a stationary time series.

1.3 Model Fitting

1.3.1 Check for Stationarity

The Augmented Dickey-Fuller (ADF) test was carried out to test the below hypothesis for stationarity. The test result is shown in figure 2.2.

H_0: The time series data is not stationary
H_1: The time series data is stationary



```
                Augmented Dickey-Fuller Test

data:  TSData1
Dickey-Fuller = -7.0257, Lag order = 14, p-value = 0.01
alternative hypothesis: stationary
```

Figure 2.2 – ADF Test for Stationarity

Here the p-value if less than α = 0.05. Therefore, at 5% level of significance, we reject the null hypothesis. Thus, we can conclude that the data is stationary.

The stationarity can be further studied using the correlogram showing the autocorrelation function. Figure 2.3 shows the autocorrelation plot. As explained before in section 2.3, the fast-decaying ACF plots suggests stationarity.

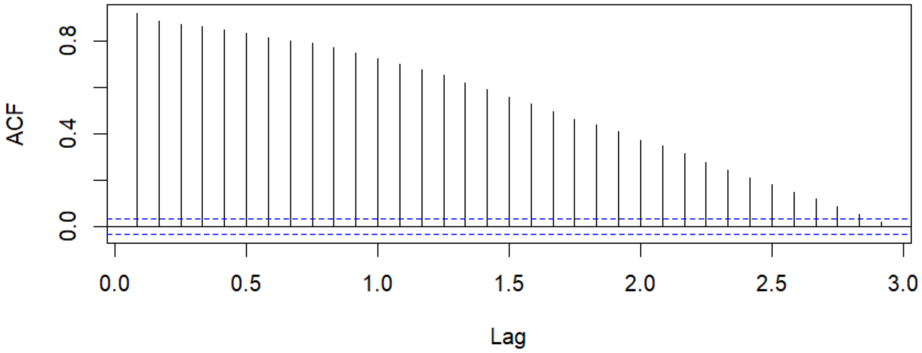


Figure 2.3 – Correlogram of Sunspots Data

Figure 2.4 shows the partial autocorrelation function. Significant peaks are observed at 1st, 2nd, 3rd and 4th lags. This suggests an AR(1), AR(2), AR(3) or AR(4) model.

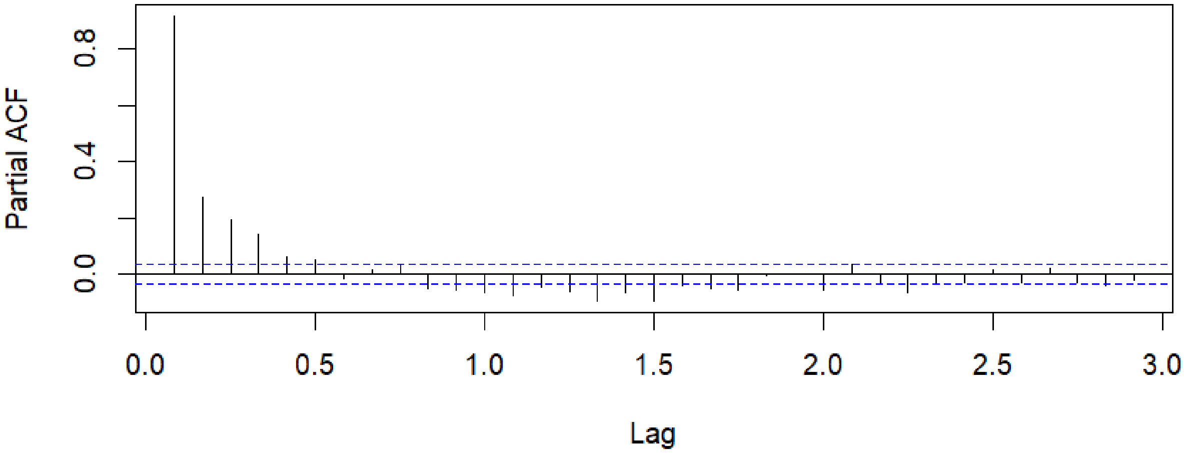
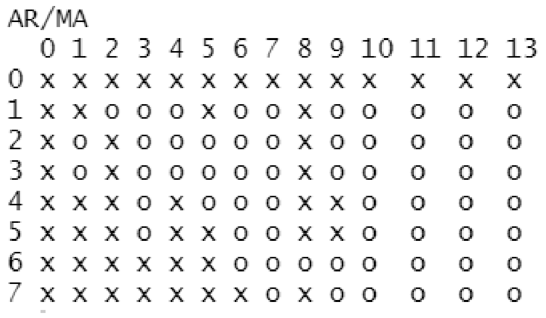


Figure 2.4 – Partial Autocorrelation Function of Sunspots Data

The EACF matrix obtained is shown in figure 2.5 below. As per the matrix AR(1) and a MA(9) might be appropriate. However, it is a bit unclear as to what model would be best suited. Hence, several models would be tested find the best fit.

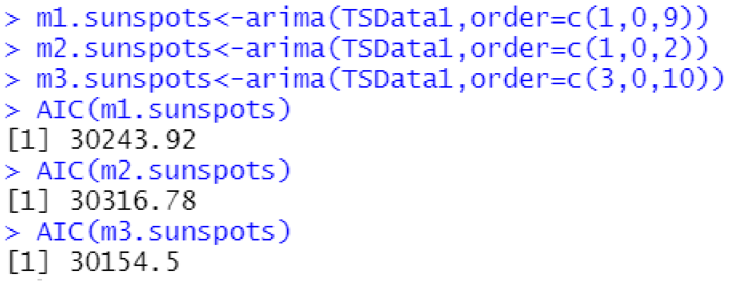


```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x o o o x o o x o o  o  o  o
2 x o x o o o o o x o o  o  o  o
3 x o x o o o o o x o o  o  o  o
4 x x x o x o o o x x o  o  o  o
5 x x x o x x o o x x o  o  o  o
6 x x x x x x o o o o o  o  o  o
7 x x x x x x x o x o o  o  o  o
```

Figure 2.5 – EACF Matrix

1.3.4 Parameter Estimates for the Model

ARMA(1,9), ARMA(1,2) and ARMA(3,10) were checked.



```
> m1.sunspots<-arima(TSData1,order=c(1,0,9))
> m2.sunspots<-arima(TSData1,order=c(1,0,2))
> m3.sunspots<-arima(TSData1,order=c(3,0,10))
> AIC(m1.sunspots)
[1] 30243.92
> AIC(m2.sunspots)
[1] 30316.78
> AIC(m3.sunspots)
[1] 30154.5
```

Figure 2.6 – ARMA Models

As per above models and AIC values, ARMA(3,10) seem to be the best model.

However, the residuals of this model must be checked for assumptions. The residuals must be a white noise.

1.3.5 Residuals Check

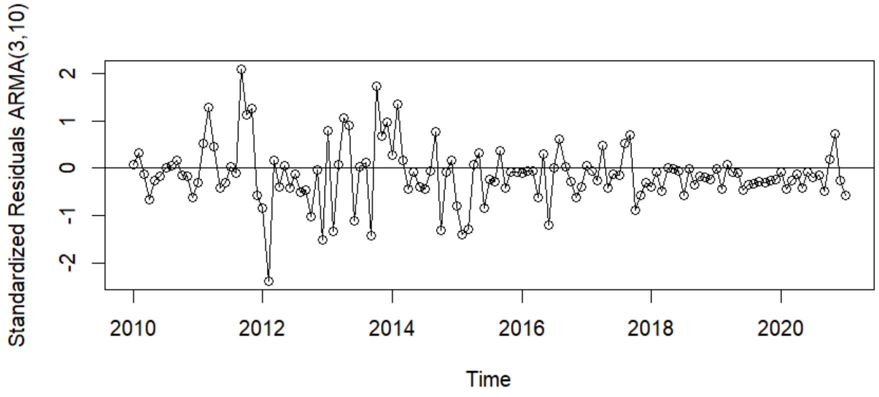


Figure 2.7 – Residuals plot of ARMA(3,10)

As per figure 2.7 the residuals look fairly alright as they are very close to zero.

The ACF of this model in figure 2.8 shows that ACF of almost all the residuals are within the interval of confidence. Hence, it appears that residuals are uncorrelated.

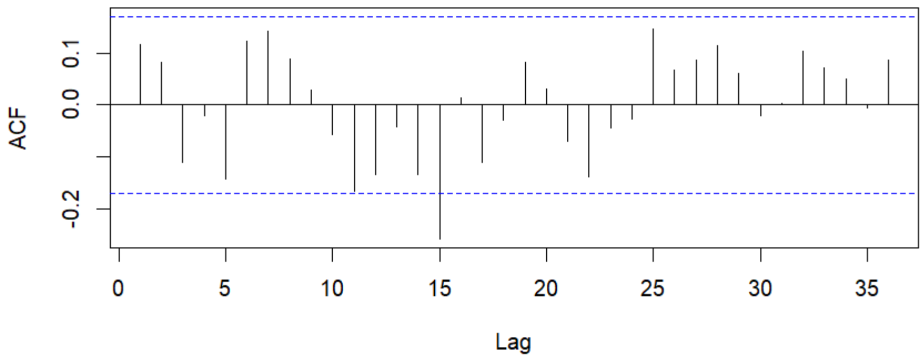
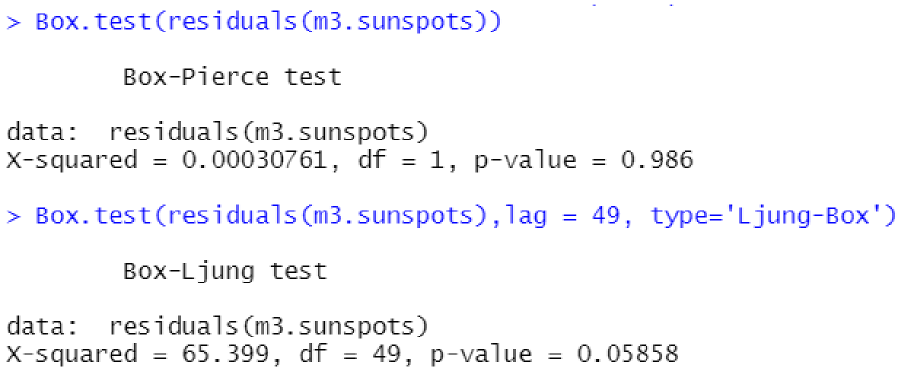


Figure 2.8 – ACF of Residuals of ARMA(3,10)

Box-Pierce test and Box-Ljung test were performed to check this further. The below hypothesis was tested.

H_o: The error terms are uncorrelated
H_1: The error terms are correlated

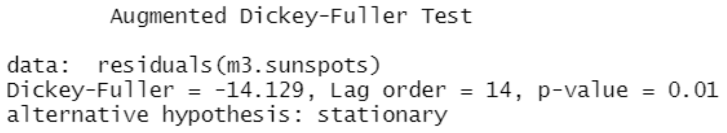


```
> Box.test(residuals(m3.sunspots))

        Box-Pierce test

data:  residuals(m3.sunspots)
X-squared = 0.00030761, df = 1, p-value = 0.986

> Box.test(residuals(m3.sunspots),lag = 49, type='Ljung-Box')

        Box-Ljung test

data:  residuals(m3.sunspots)
X-squared = 65.399, df = 49, p-value = 0.05858
```

Figure 2.9- Box-Pierce Test and Box-LjungTest

As per both figure 2.9 above the p-values are greater than α=0.05, indicating that the null hypothesis cannot be rejected. Hence, at 5% level of significance we have enough evidence to say that the error terms are uncorrelated. Thus, the residuals do not depend on time.

Thereafter, the ADF test was also performed to check the stationarity of the residuals. Figure 2.10 shows the output. Here, the p-value is less than α=0.05. Therefore, at 5% level of significance we reject the null hypothesis that the residuals are not stationary. This shows that the residuals are stationary.

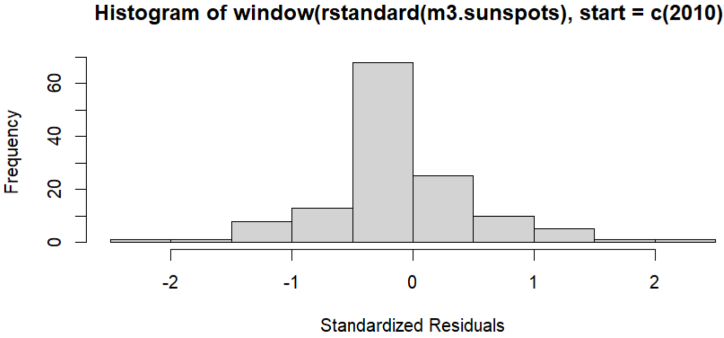


```
        Augmented Dickey-Fuller Test

data:  residuals(m3.sunspots)
Dickey-Fuller = -14.129, Lag order = 14, p-value = 0.01
alternative hypothesis: stationary
```

Figure 2.10- ADF Test for Residuals

Therefore, we can conclude that the residuals follow a white noise.

The residuals must also follow a normal distribution. Thus, the following plots were checked.



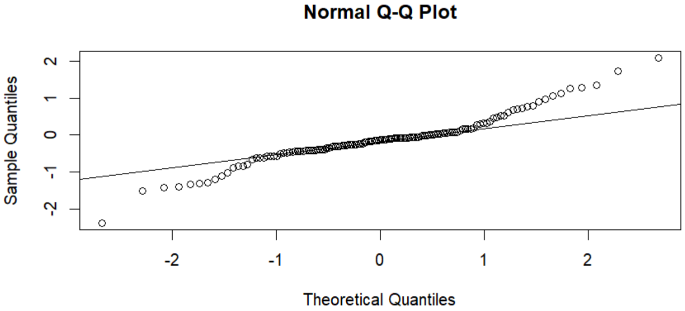
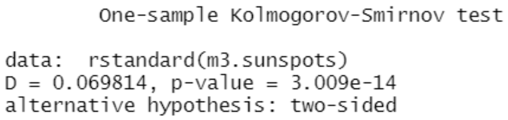


Figure 2.11 - Histogram of Residuals

Figure 2.12 – Q-Q Plot

The histogram of residuals (Figure 2.11) and Q-Q plot (Figure 2.12) both show that the residuals may be following a normal distribution.



```
	One-sample Kolmogorov-Smirnov test

data:  rstandard(m3.sunspots)
D = 0.069814, p-value = 3.009e-14
alternative hypothesis: two-sided
```

Figure 2.13 Kolmogorov-Smirnov Test

Kolmogorov-Smirnov test was done to further check the normality, and its output in Figure 2.13 shows that p-value is less than α = 0.05, indicating the null hypothesis that residuals follow a normal distribution is rejected at 5% level of significance. Therefore, the residuals are not following a perfectly normal distribution.

2.4 Forecasting

The forecasts plot for the next 24 months is shown in figure 2.14. It does not appear to be well in line with the pattern in the previous months. Hence, by first looks the forecasts may not be 100% accurate.

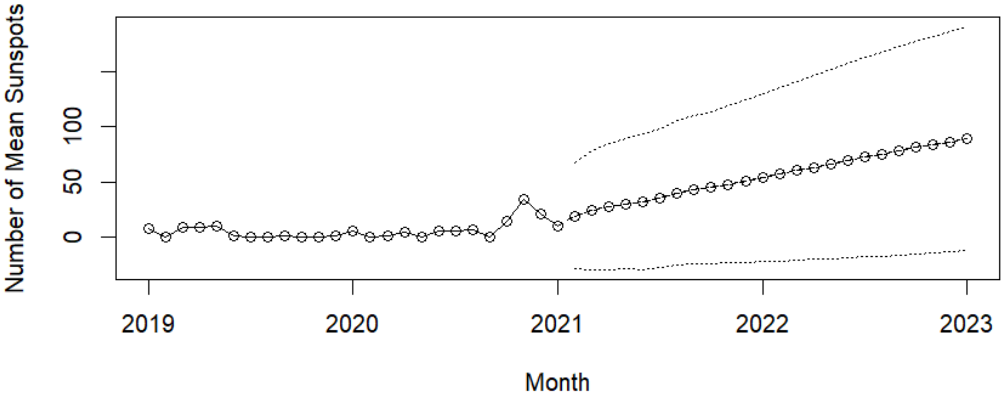
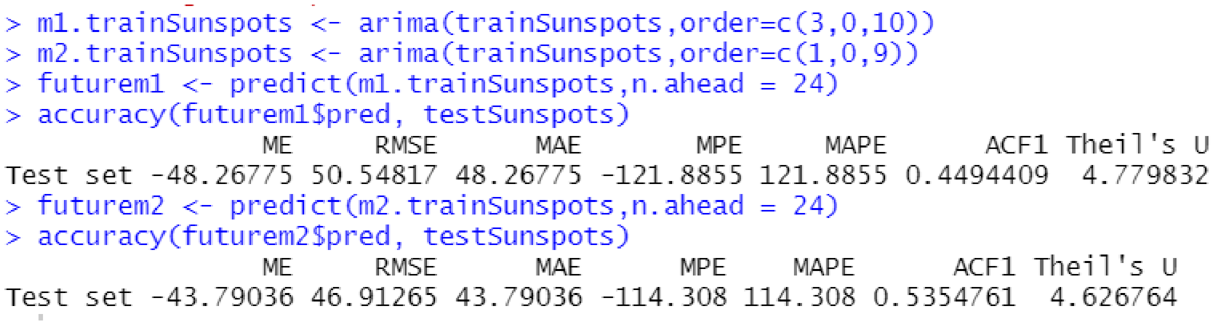


Figure 2.14 Forecasts for next 24 months

The data were then split into train and test sets.

Forecast accuracy was checked for ARMA(3,10) model and ARMA(1,9) model by comparing the predictions with the test data set. The below figure 2.15 gives the accuracy output.



```
> m1.trainSunspots <- arima(trainSunspots,order=c(3,0,10))
> m2.trainSunspots <- arima(trainSunspots,order=c(1,0,9))
> futurem1 <- predict(m1.trainSunspots,n.ahead = 24)
> accuracy(futurem1$pred, testSunspots)
                ME     RMSE      MAE       MPE     MAPE      ACF1 Theil's U
Test set -48.26775 50.54817 48.26775 -121.8855 121.8855 0.4494409  4.779832
> futurem2 <- predict(m2.trainSunspots,n.ahead = 24)
> accuracy(futurem2$pred, testSunspots)
                ME     RMSE      MAE      MPE    MAPE      ACF1 Theil's U
Test set -43.79036 46.91265 43.79036 -114.308 114.308 0.5354761  4.626764
```

Figure 2.15 Forecasts Accuracy

As per the above output, the forecasts of the ARMA(1,9) model seem to have lesser mean squared error and root mean squared error than ARMA(3,10). Hence, it would be better to use the model ARMA(1,9) for forecasting.

The models could be further explored to find a better fit.

<u>References</u>

OS visits to UK:All visits Thousands-NSA (2020) Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/timeseries/gmaa/ott (Accessed: 1 Mar 2022).

Kaggle (no date) Sunspots. Available at: https://www.kaggle.com/robervalt/sunspots (Accessed: 1 Mar 2022).