TIME SERIES ANALYSIS AND FORECASTING OF MONTHLY MEAN SUNSPOTS NUMBER

MATH1318 FINAL PROJECT

Kaushik Sunil Anagarkar - S3827495
Pallavi Bhimte - S3758167
Parth Deshmukh - S3825055
Pooja Mallinath - S3820735
Subbiah Soundarapandian - S3825012
Sudershan Ravi - S3829895

Contents

INTRODUCTION:

- This report examines the study of monthly mean sunspot numbers using time series analysis methods.
- We fit the required time series regression models to our time series data and decide on the best data based on the residual analysis.
- Finally, we Predict monthly mean sunspot counts for the next ten years and forecast for the same years.

DATA DESCRIPTION:

- The sunspot dataset is sourced from Kaggle[1].
- The number of sunspots changes every 11 years or so, depending on the solar cycle. The dataset spans the years 1749 to 2021.
- After then, it is turned into a time series object. Monthly Mean Total Sunspot Number is selected for analysis.

DATA EXPLORATION:

Time Series Plot:

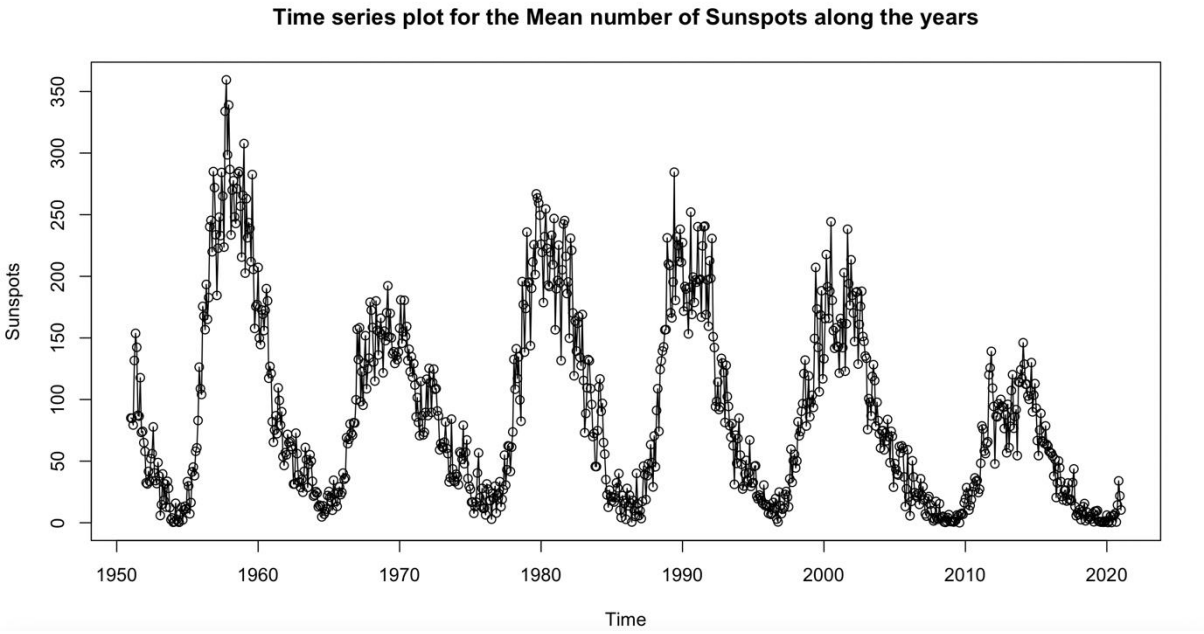


Figure 1 Time Series Plot for the mean number of sunspots along the years

As the trend is same throughout the years, we filter the data from 1951 to reduce the complexity of our analysis. After filtering, let's have a look at the time series plot and tell about 5 talking points of a time series plot.

Following points can be inferred by looking at the above time series plot:

1. Seasonality: A repetitive pattern is observed. Hence seasonality is present.

2. Trend: Mild trend with non-stationarity found.
3. Behavior: Mostly Moving Average with slight Auto Regressive behavior.
4. Intervention Point: There is no intervention point.
5. Change of Variance: Mild significance change of variance in the plot could be seen.

ACF & PACF Plot:

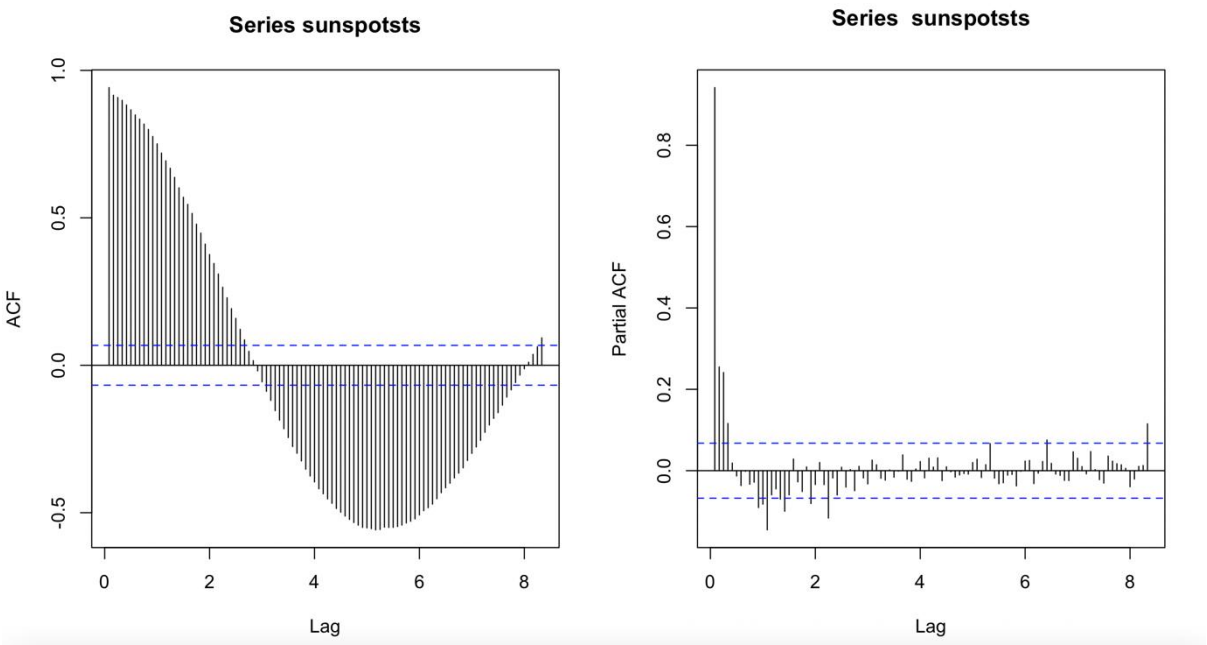


Figure 2 ACF and PACF plot of time series

ACF- After seeing the correlation between time series and its first lagged value, a decaying wave like pattern was observed indicating that time series is non-stationary and seasonal in nature.

PACF- Considering the partial correlation between the time series and its lags, there were four significant lags observed that were above the blue-dashed threshold line. It can be inferred that time series is non-stationary.

Stationarity Test:

```
> adf.test(sunspotsts)

        Augmented Dickey-Fuller Test

data:  sunspotsts
Dickey-Fuller = -2.8096, Lag order = 9, p-value = 0.2356
alternative hypothesis: stationary
```

With the help of Augmented Dickey Fuller t-statistic test, the obtained p-value was 0.236. This implies that the trend line has a unit root since the p-value was found to be large. The null hypothesis states that data is non-stationary. We do not reject the null-hypothesis as the p-value was more than the level of significance (above 0.05).

ANALYSIS:

The code used in carrying out the statistical analysis of the given time series data is included in APPENDIX A.

MODEL SPECIFICATION:

Since the time series had seasonality, we go for SARIMA models to predict and forecast. We opt for the following approach for the SARIMA model,

Residual Approach:

Residual approach is selected over the classical approach as the former is simpler and clearer while being similar to the latter. The residual approach is easier to implement and to understand which makes it an obvious choice.

Residuals consists of the data not captured by the model hence, giving us an insight about the fit and performance of the model. To achieve this, the residuals can be examined to check for leftovers after model fitting and this process can be repeated to refine the model until all the data and its autocorrelation between them, is captured.

Autocorrelation Function and Partial Autocorrelation Function (ACF and PACF):

Next, the ACF and PACF plots for the average number of sunspots in the time series are generated to examine the seasonality and the lags present as the frequency is greater than 1.

It is evident from the figure 2, that there is a decaying pattern along with seasonal lags. This calls for the requirement of differencing the time series to correct the non-stationarity. Hence, we start our residuals approach by applying first non-seasonal differencing (d) to our data and examine further. We also pass the parameter $\lambda = 0.47$ to Box-Cox transform the time series data to get a stable variance throughout the data. Now let us have a look at ACF and PACF of this fit's residuals.

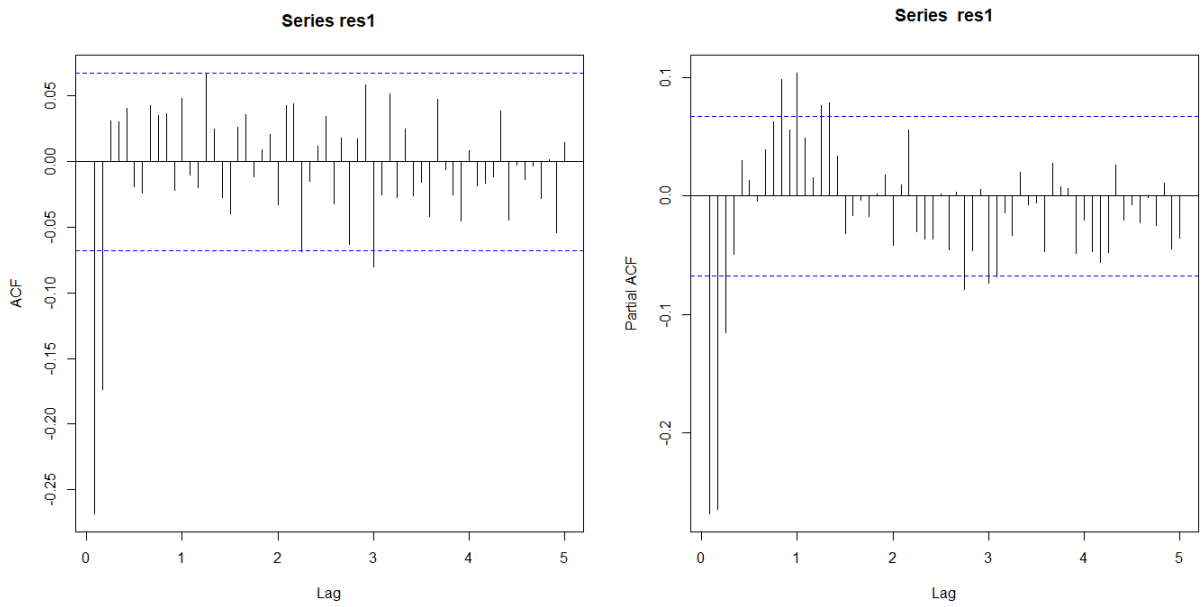


Figure 3 ACF AND PACF OF RESIDUALS - FIT 1

The first set of model specification can be assumed to be **SARIMA (0,1,0)x(0,0,0)(12)**, the s value is designated to be 12 as it is an annual time series consisting of 12 months.
As mentioned earlier, the first differencing was performed to achieve stationarity following which the ACF and PCF plots were generated to check for decaying pattern and seasonality. Since, it is confirmed that the time series is now stationary the AR, MA (p,d,q) and seasonal AR,MA (P,D,Q) values can be determined from the plots.

We observe the ACF plot first, for the q and seasonal Q values. Meanwhile the PACF plot provides us the p and seasonal P values. To find the p and q values the number of significant bars before the first lag. And for the seasonal P and Q values we observe the significance bars corresponding to the seasonal lags.

From the above ACF plot, we can determine that the q and seasonal Q values as 2 and 0 respectively while, on observing the PACF plot the p and seasonal P values could be estimated to be 3 and 1, respectively. So, the second model that can be specified is **SARIMA (3,1,2)x(1,0,0)(12).**

So, according to the residual approach strategy the ACF and PACF plots are examined, and it is discovered that there is a seasonal decaying pattern resulting in significant correlation present which calls for further model fitting iteration. Hence, the second iteration of model fitting performed with the use of the model specification procured in the previous step and the plots observed.

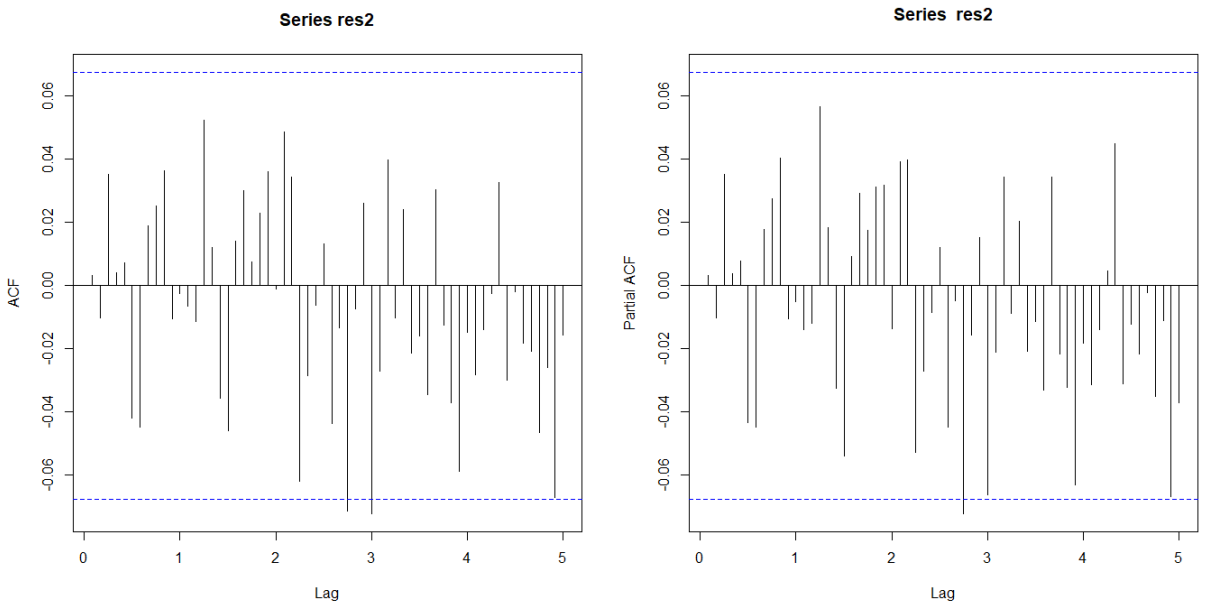
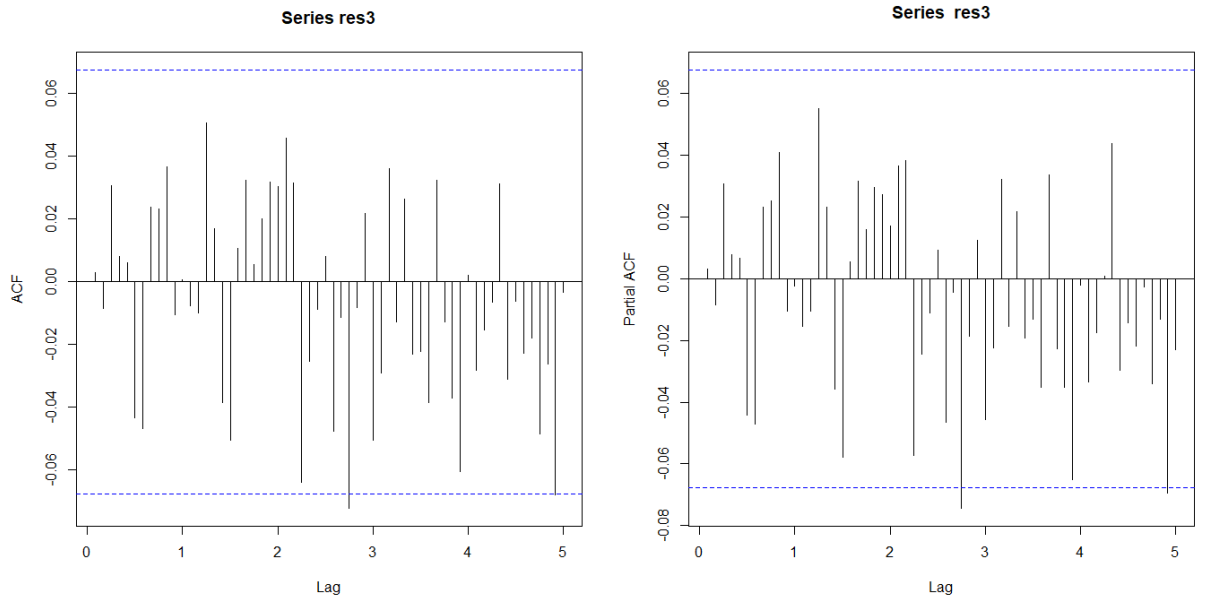


Figure 4 ACF AND PACF OF RESIDUALS - FIT 2

The above plots evidently provide the seasonal P and Q values as 1 and 2 respectively as there is a significant bar at first seasonal lag in the PACF and at second seasonal lag in ACF plot. Hence, the third set of model specification can be assumed as **SARIMA (3,1,2)x(1,0,2)(12).**

Subsequently, the third iteration of model fitting is performed using the above specified model is fit and the residuals are plotted for examination.



The ACF and PACF subsequently generated reveals the third set of seasonal P and Q values. And from the above output it can confidently assumed that there are no significant ordinary or seasonal lag values. Hence, it is safe to say that there is no seasonal significant autocorrelation left in the residuals.

Extended Autocorrelation Function (EACF)

Following working with the seasonal orders, an EACF plot is generated to determine the ordinary lags. The EACF matrix provides the p and q values from the x and y axes. The best model could be determined by locating the top left '0' in the matrix and the adjacent zeros provide the other sets of (p,d,q) values. To achieve this, the residuals that corrected the seasonal lags is used.

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

Figure 5 EACF - FIT 3

From the above matrix, it can be determined that the (p,d,q) as (0,1,0) , (0,1,1), (1,1,1) are the sets of ordinary orders. Hence, the resultant set of model specifications are **SARIMA (0,1,0)x(1,0,2)(12), SARIMA (0,1,1)x(1,0,2)(12) and SARIMA (1,1,1)x(1,0,2)(12).** These model's efficacy can be confirmed in model fitting and residual analysis section.

Bayesian Information Criterion (BIC)

Similarly, a BIC plot is another visualization to determine the non-seasonal parameters (p,d,q).

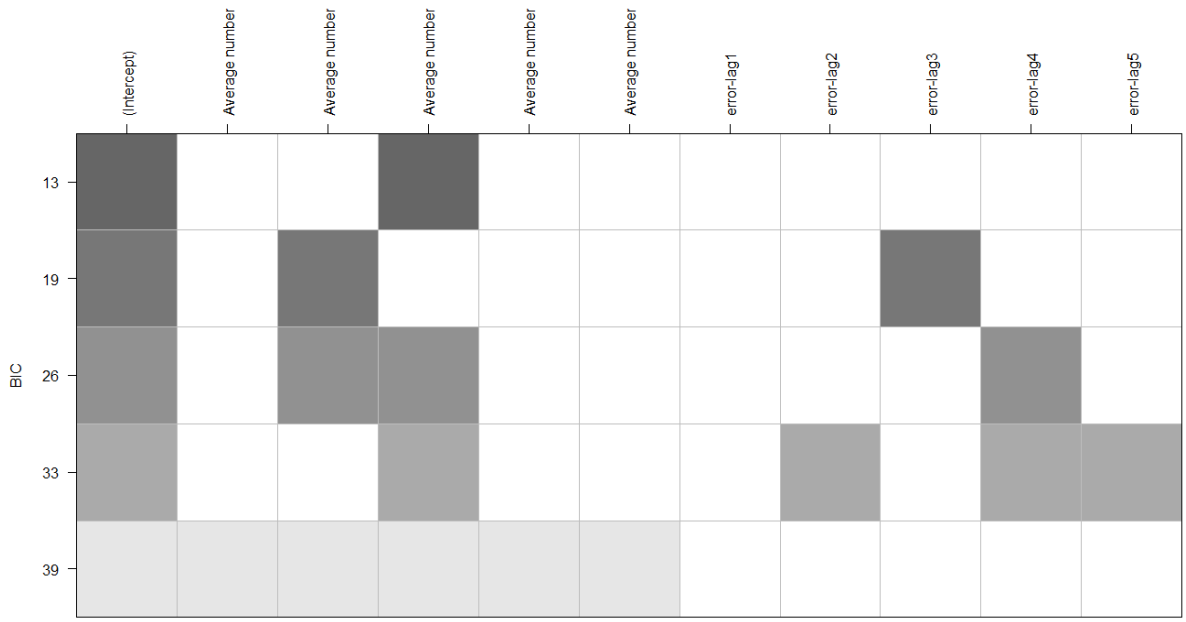


Figure 6 BIC - FIT 3

From the above plot, it is safe to assume that the (p,d,q) values to be (3,1,0) as the 3rd p value has the minimum BIC value though it is not strongly supported while there are no

significant bars . Therefore, the final set of model specification would be **SARIMA (3,1,0)x(1,0,2)(12).**

To summarize, the possible sets of models are as follows:
SARIMA (0,1,0)x(0,0,0)(12), SARIMA (3,1,2)x(1,0,0)(12), SARIMA (3,1,2)x(1,0,2)(12), SARIMA (0,1,0)x(1,0,2)(12), SARIMA (0,1,1)x(1,0,2)(12), SARIMA (1,1,1)x(1,0,2)(12) and SARIMA (3,1,0)x(1,0,2)(12).

MODEL FITTING

As mentioned earlier the data shows some seasonality so we will escalate our analysis using SARIMA model. The Residuals approach was used to calculate the EACF and BIC of the residuals in order to decide the non-seasonal parameters (p, d, q). The seasonal parameters for the model were chosen based of their residuals for which (1, 0, 2) model was the best as it did not contain any significant seasonal lags.

Using these seasonal and non-seasonal parameters along with CSS-ML method we obtained seven models for SARIMA.

SARIMA$(0,1,0)x(0,0,0)_{12}$, SARIMA $(3,1,2)x(1,0,0)_{12}$, SARIMA$(3,1,2)x(1,0,2)_{12}$, SARIMA$(0,1,0)x(1,0,2)_{12}$, SARIMA$(0,1,1)x(1,0,2)_{12}$, SARIMA$(1,1,1)x(1,0,2)_{12}$, SARIMA$(3,1,0)x(1,0,2)_{12}$

Now let us fit each of the above models and decide on the best model based on its auto-regressive and moving average parameters.

1] MODEL 1 - *SARIMA$(0,1,0)x(0,0,0)_{12}$*

SARIMA Model 1 has only first differencing with no p and q values. Also, no seasonal parameters are assigned. Hence, we cannot obtain any results for the specified seasonal and non-seasonal parameter and will proceed with other models.

2] MODEL 2 - *SARIMA $(3,1,2)x(1,0,0)_{12}$*

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   1.253727   0.070258  17.8446 < 2.2e-16 ***
ar2  -0.386006   0.059421  -6.4961 8.244e-11 ***
ar3   0.079454   0.046678   1.7022   0.08872 .
ma1  -1.666399   0.063723 -26.1508 < 2.2e-16 ***
ma2   0.717921   0.060465  11.8733 < 2.2e-16 ***
sar1  0.043405   0.037126   1.1691   0.24235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7 FIT 2 –Seasonal and Non-Seasonal MA AND AR PARAMETERS

SARIMA Model 2 has three Auto-Regressive parameteres out of which first two (ar1) and (ar2) are highly significant. Similary both the Moving-Average parameters (ma1) and (ma2)

are highly significant whereas other parameters (ar1) and (sar1) are unsignificant. But, still it can be a best model.

3] MODEL 3 - *SARIMA(3,1,2)x(1,0,2)$_{12}$*

```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   1.249609   0.071191  17.5528  < 2e-16 ***
ar2  -0.384207   0.059544  -6.4525  1.1e-10 ***
ar3   0.081123   0.046787   1.7339  0.08294 .
ma1  -1.662851   0.064584 -25.7472  < 2e-16 ***
ma2   0.713548   0.061249  11.6499  < 2e-16 ***
sar1  0.767520   0.324161   2.3677  0.01790 *
sma1 -0.728049   0.321983  -2.2611  0.02375 *
sma2 -0.062577   0.035454  -1.7650  0.07756 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8 FIT 3 –Seasonal and Non-Seasonal MA AND AR PARAMETERS

SARIMA Model 3 has three Auto-Regressive parameteres out of which first two (ar1) and (ar2) are highly significant. Moreover, both the Moving-average (ma1) and (ma2) parameters are highly significant, in addition to that Seasonal Auto-Regressive (sar1) and Seasonal Moving-average (sma1) both are mildy significant and hence can be a best model.

4] MODEL 4 - *SARIMA(0,1,0)x(1,0,2)$_{12}$*

```
z test of coefficients:

      Estimate Std. Error z value Pr(>|z|)
sar1 -0.680329   0.477908 -1.4236   0.1546
sma1  0.734745   0.475730  1.5445   0.1225
sma2  0.048887   0.040593  1.2043   0.2285
```

Figure 9 FIT 4 –Seasonal and Non-Seasonal MA AND AR PARAMETERS

SARIMA Model 4 has no Auto-Regressive parameteres, and do not have any significant parameters among Seasonal Auto-Regressive (sar1) , Seasonal Moving-average (sma1) and Seasonal Moving-average (sma2). Hence, it cannot be a best model.

5] MODEL 5 - *SARIMA(0,1,1)x(1,0,2)$_{12}$*

```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ma1  -0.456867   0.034244 -13.3417  < 2e-16 ***
sar1 -0.515688   0.321896  -1.6020  0.10915
sma1  0.591622   0.320085   1.8483  0.06455 .
sma2  0.074200   0.041258   1.7984  0.07211 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10 FIT 5 –Seasonal and Non-Seasonal MA AND AR PARAMETERS

SARIMA Model 5 has no Auto-Regressive parameteres, and only Moving-average (ma1) is a highly significant parameter. Other parameters Seasonal Auto-Regressive (sar1) , Seasonal Moving-average (sma1) and Seasonal Moving-average (sma2) are unsignificant and hence it cannot be a best model.

6] MODEL 6 - *SARIMA(1,1,1)x(1,0,2)$_{12}$*

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.205738   0.058258   3.5315 0.0004132 ***
ma1  -0.605437   0.044065 -13.7395 < 2.2e-16 ***
sar1  0.802730   0.378945   2.1183 0.0341472 *
sma1 -0.727449   0.372833  -1.9511 0.0510406 .
sma2 -0.104464   0.034160  -3.0580 0.0022278 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  .
```

Figure 11 FIT 6 –Seasonal and Non-Seasonal MA AND AR PARAMETERS

SARIMA Model 6 has one Auto-Regressive parametere (ar1) which is highly significant along with Moving-avergae (ma1). Moreover, both the Seasonal Auto-Regressive (sar1) and Seasonal Moving-average (sma2) parameters are midly significant and Seasonal Moving-average (sma1) is unsignificant. Hence, cannot be a best model.

7] MODEL 7 - *SARIMA(3,1,0)x(1,0,2)$_{12}$*

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1  -0.376145   0.034403 -10.9337 < 2.2e-16 ***
ar2  -0.313221   0.035231  -8.8904 < 2.2e-16 ***
ar3  -0.133281   0.035227  -3.7835 0.0001546 ***
sar1  0.917912   0.070963  12.9351 < 2.2e-16 ***
sma1 -0.852958   0.078234 -10.9027 < 2.2e-16 ***
sma2 -0.092651   0.034974  -2.6491 0.0080695 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12 FIT 7 –Seasonal and Non-Seasonal MA AND AR PARAMETERS

SARIMA Model 7 has three Auto-Regressive parametere (ar1) , (ar2) and (ar3) are highly significant. Moreover, Seasonal Auto-Regressive (sar1) and Seasonal Moving-average (sma1) are also highly significant, whereas Seasonal Moving-average (sma2) is midly significant. Hence, it can be a best model.

After, carefully analyzing all the seven models we found that *SARIMA(3,1,2)x(1,0,2)*$_{12}$ was the best model among all with seasonal parameters (1 , 0 , 2) and non-seasonal parameters (3 1 , 2).

Residual Analysis

- If the model is correctly specified and the parameter estimates are reasonably close to the true values, then the residuals should have nearly the properties of white noise. They should behave roughly like independent, identically distributed normal variables with zero means and common standard deviations. Deviations from these properties can help us discover a more appropriate model.

Now let us have a look at the residual analysis of each of the fits obtained previously.

Plots of the Residuals:

1. Residual analysis of SARIMA (0,1,0)x(0,0,0)$_{(12)}$ model

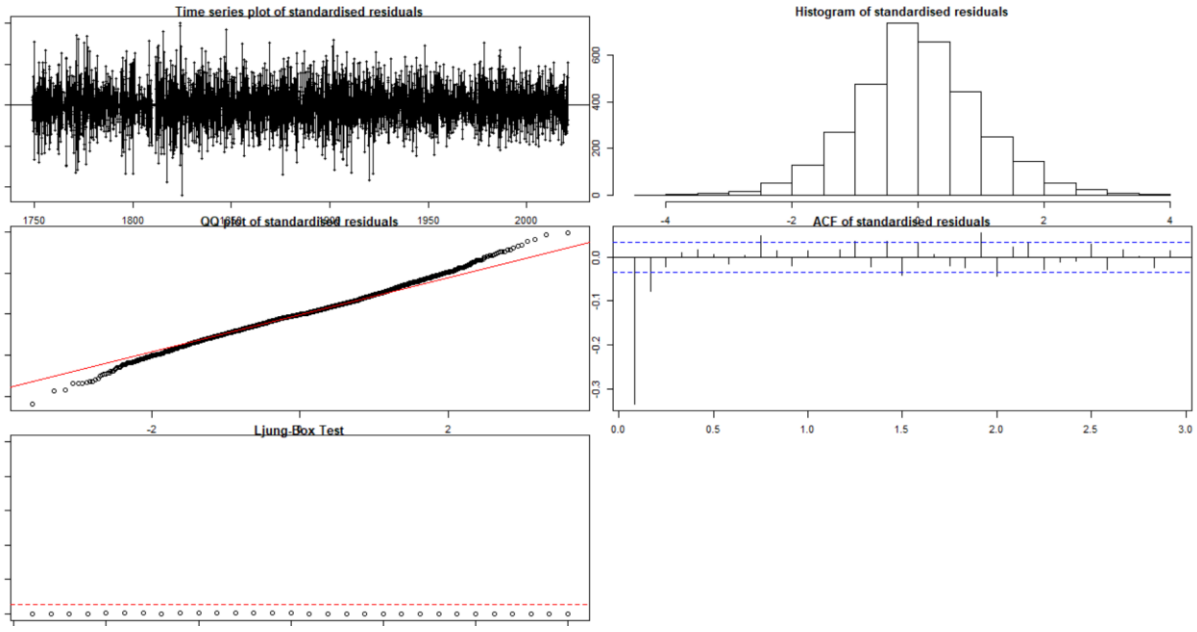


Figure 13 Residual Analysis Plots - FIT 1

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.99542, p-value = 1.624e-08
```

- If the model is adequate, we expect the plot to suggest a rectangular scatter around a zero-horizontal level with no trends whatsoever.

- Time series plot of the residuals shows that it does not capture any trend and hence it becomes a white noise series.
- Histogram: The Histogram of the Residual can be used to check whether the variance is normally distributed. A symmetric bell-shaped histogram which is evenly distributed around zero indicates that the normality assumption is likely to be true. Here, the histogram indicates that random error is almost normally distributed.
- Q-Q plot: We apply quantile-quantile plots for the analysis of normality of residuals. Because majority of points including the extreme ones closely follow the straight line, we cannot reject normality of the error terms in this model.
- Shapiro-Wilk test: With P-value is less than 0.01, so we rejected the null hypothesis that the stochastic component of this model is normally distributed.
- ACF plot: To check on the independence of the noise terms in the model, we consider the sample autocorrelation function of the residuals. Few significant lags that confirm there are some autocorrelations left in the residuals.
- The Ljung-Box Test: The Ljung-Box test provides an overall test for looking at residual correlations. For example, it may be that most of the residual autocorrelations are moderate, some even close to their critical values, but, together, they seem excessive. Here, all points are below red dashed line, we have evidence to reject the null hypothesis that the error terms are uncorrelated.
- Overall, based on residual analysis above, normality did not hold for SARIMA(0,0,0)x(0,1,0)(12) model**.** The model is not a good fit.

2. Residual analysis of SARIMA $(3,1,2)x(1,0,0)_{(12)}$ model

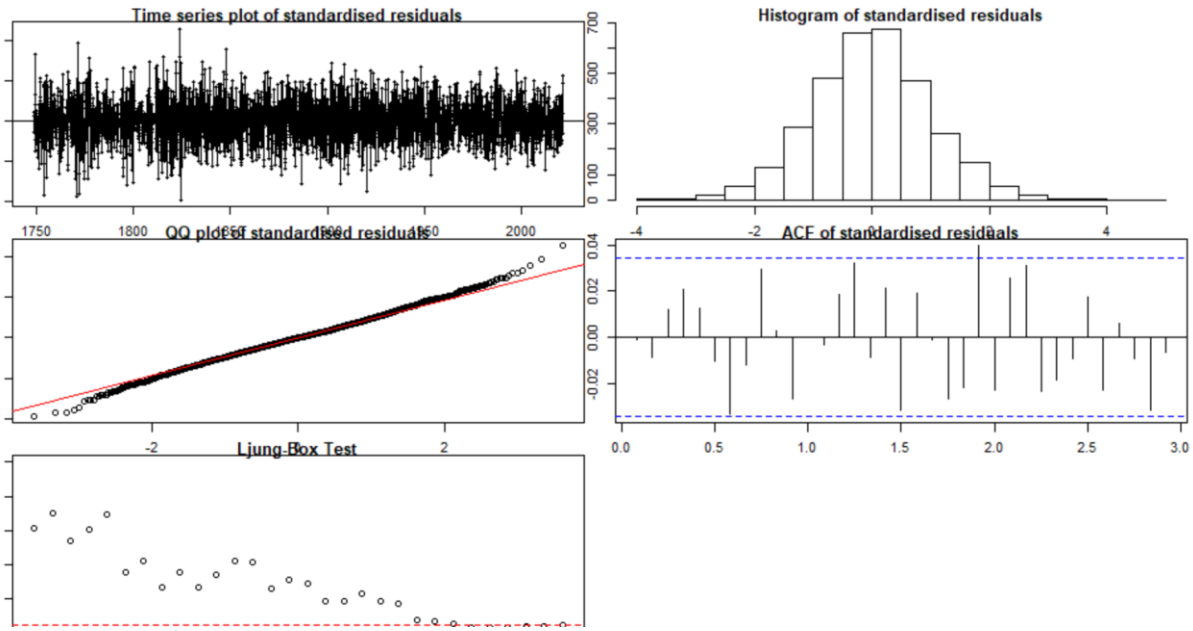


Figure 14 Residual Analysis Plots - FIT 2

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.99767, p-value = 7.8e-05
```

- Time series plot of the residuals shows that it does not capture any trend and hence it becomes a white noise series.
- Histogram of the Residuals showing that the deviation is normally distributed.
- Q-Q plot: We apply quantile-quantile plots for the analysis of normality of residuals. Because majority of points including the extreme ones closely follow the straight line, we cannot reject normality of the error terms in this model.
- Shapiro-Wilk test: With P-value is less than 0.01, so we rejected the null hypothesis that the stochastic component of this model is normally distributed.
- ACF plot: Few significant lags that confirm there are some autocorrelations left in the residuals.
- The Ljung-Box Test: Most of the points are above red dashed line, we have no evidence to reject the null hypothesis that the error terms are uncorrelated.
- Overall, based on residual analysis above, normality did not hold for SARIMA(3,1,2)x(1,0,0)$_{(12)}$ model**.** The model is not a good fit.

3. Residual analysis of SARIMA (3,1,2)x(1,0,2)$_{(12)}$ model

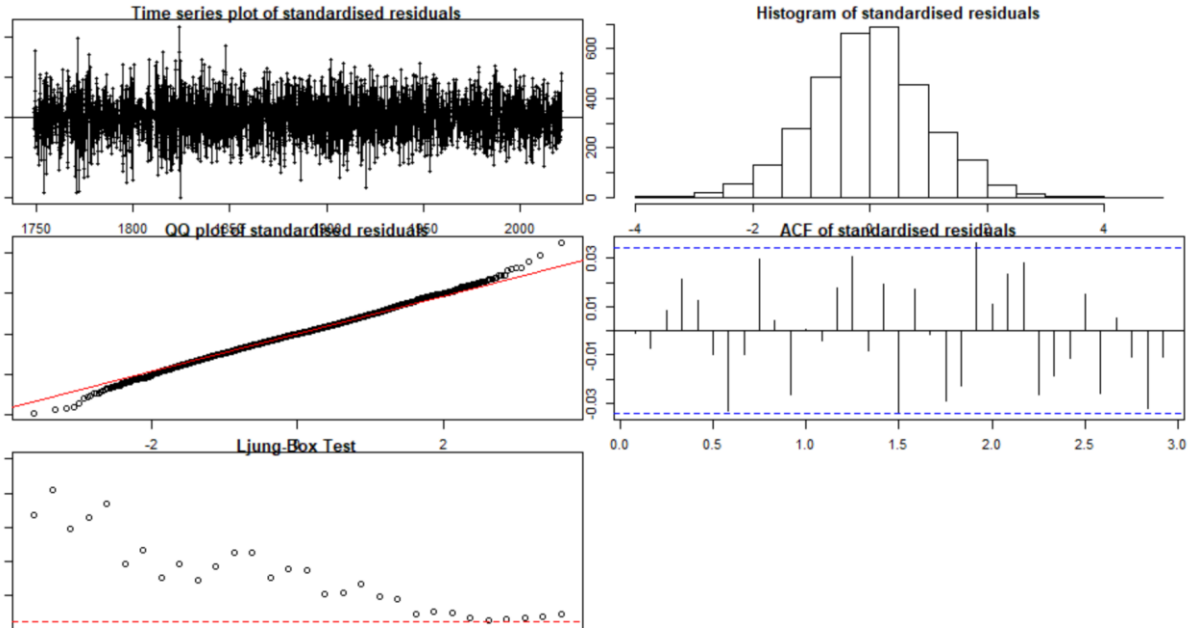


Figure 15 Residual Analysis Plots - FIT 3

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.99767, p-value = 8.004e-05
```

- Time series plot of the residuals shows that it does not capture any trend and hence it becomes a white noise series.
- Histogram of the Residuals showing that the deviation is normally distributed.

- Q-Q plot: QQ Plot of the standardised residuals shows that it is normally distributed between the desired range of -3 and +3.
- Shapiro-Wilk test: With P-value is less than 0.01, so we rejected the null hypothesis that the stochastic component of this model is normally distributed.
- ACF plot: The ACF plot of residuals shows that, it does not capture any significant autocorrelation which suggest that model is good enough to capture all the significant information.
- Ljung Box test has the null hypothesis that model has captured the significant autocorrelations. Here as we could see, the p-value is greater than the significance level in all the lags and hence it shows that the model is a good fit.

4. Residual analysis of SARIMA $(0,1,0)x(1,0,2)_{(12)}$ model

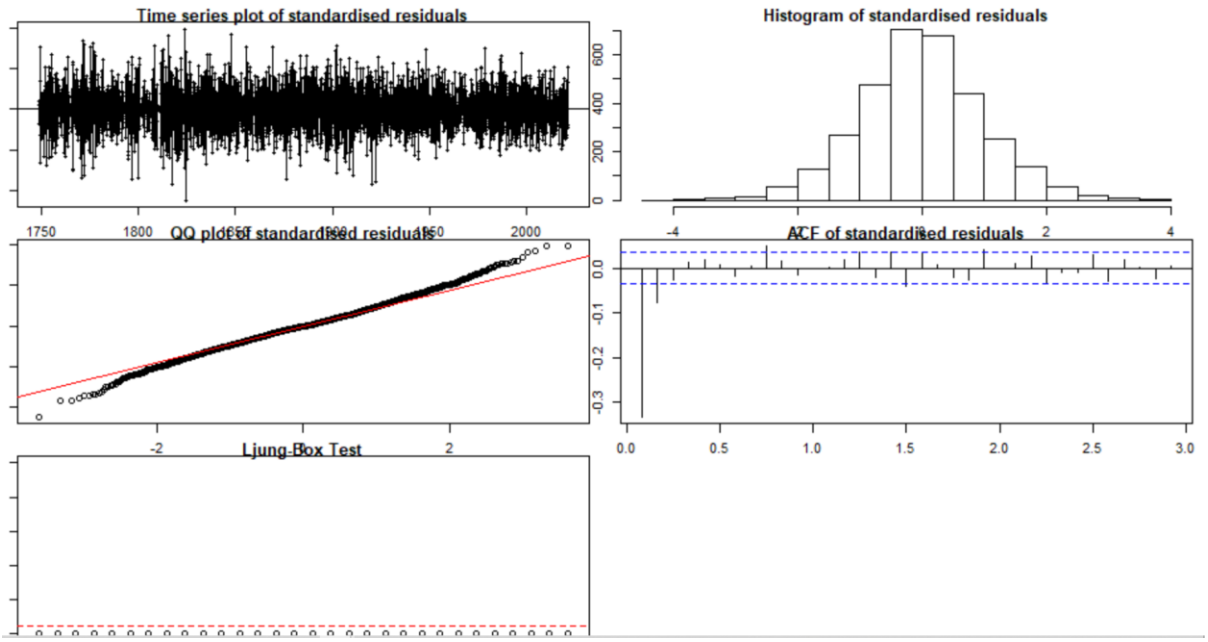


Figure 16 Residual Analysis Plots - FIT 4

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.99551, p-value = 2.142e-08
```

- Time series plot of the residuals shows that it does not capture any trend and hence it becomes a white noise series.
- Histogram of the Residuals showing that the deviation is normally distributed.
- Q-Q plot: We apply quantile-quantile plots for the analysis of normality of residuals. Because majority of points including the extreme ones closely follow the straight line, we cannot reject normality of the error terms in this model.
- Shapiro-Wilk test: With P-value is less than 0.01, so we rejected the null hypothesis that the stochastic component of this model is normally distributed.

- ACF plot: Few significant lags that confirm there are some autocorrelations left in the residuals.
- The Ljung-Box Test: Most of the points are above red dashed line, we have no evidence to reject the null hypothesis that the error terms are uncorrelated.

Overall, based on residual analysis above, normality did not hold for SARIMA(1,0,2)x(0,1,0)(12) model**.** The model is not very successful to be a good fit.

5. Residual analysis of SARIMA $(0,1,1)x(1,0,2)_{(12)}$ model

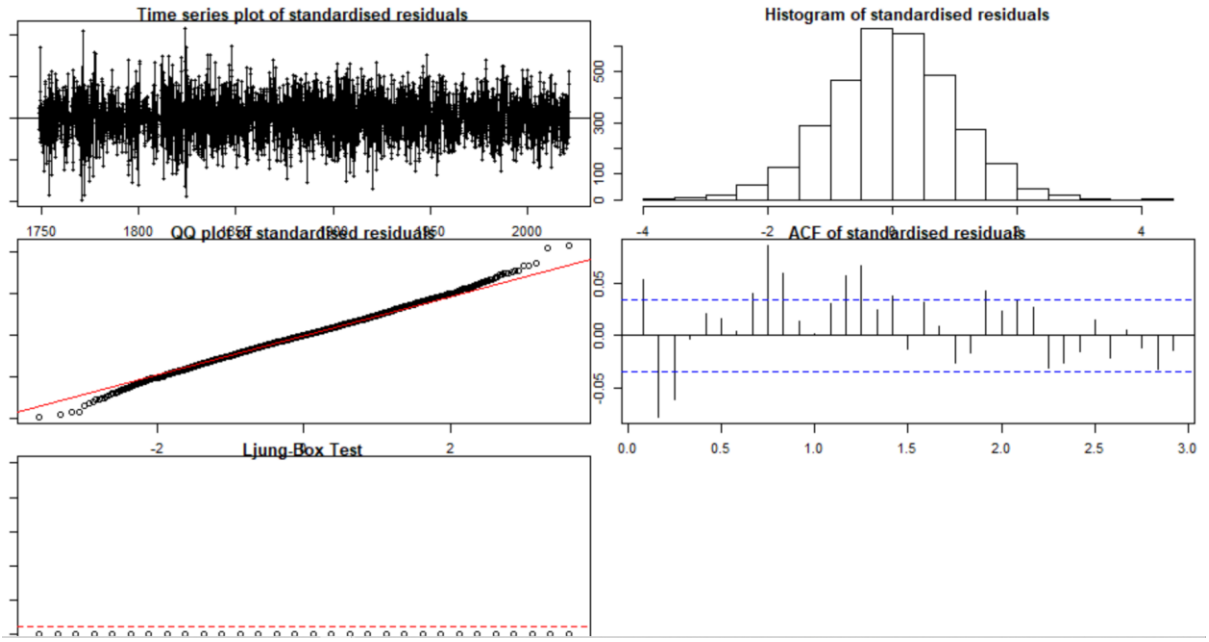


Figure 17 Residual Analysis Plots - FIT 5

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.99795, p-value = 0.0002831
```

- Time series plot of the residuals shows that it does not capture any trend and hence it becomes a white noise series.
- Histogram of the Residuals showing that the deviation is normally distributed.
- Q-Q plot: We apply quantile-quantile plots for the analysis of normality of residuals. Because majority of points including the extreme ones closely follow the straight line, we cannot reject normality of the error terms in this model.
- Shapiro-Wilk test: With P-value is less than 0.01, so we rejected the null hypothesis that the stochastic component of this model is normally distributed.
- ACF plot: Few significant lags that confirm there are some autocorrelations left in the residuals.
- The Ljung-Box Test: All points are below red dashed line, we have evidence to reject the null hypothesis that the error terms are uncorrelated.

Overall, based on residual analysis above, normality did not hold for SARIMA $(0,1,1)x(1,0,2)_{(12)}$ model**.** The model is not a good fit.

6. Residual analysis of SARIMA $(1,1,1)x(1,0,2)_{(12)}$ model

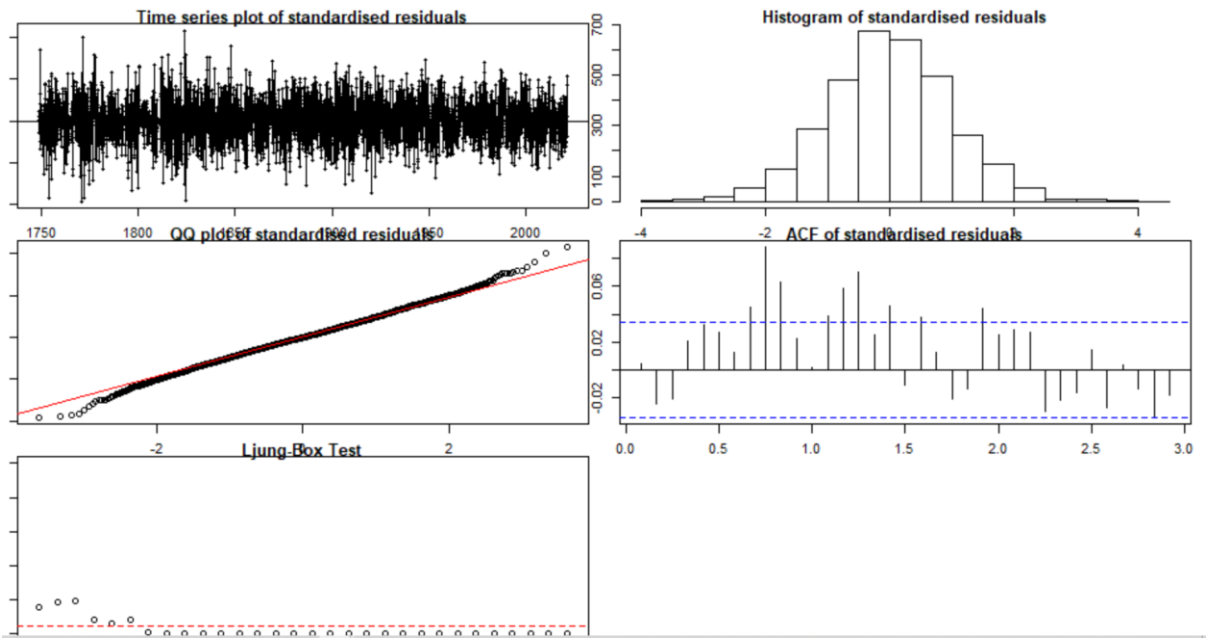


Figure 18 Residual Analysis Plots - FIT 6

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.99794, p-value = 0.0002705
```

- Time series plot of the residuals shows that it does not capture any trend and hence it becomes a white noise series.
- Histogram of the Residuals showing that the deviation is normally distributed.
- Q-Q plot: We apply quantile-quantile plots for the analysis of normality of residuals. Because majority of points including the extreme ones closely follow the straight line, we cannot reject normality of the error terms in this model.
- Shapiro-Wilk test: With P-value is less than 0.01, so we rejected the null hypothesis that the stochastic component of this model is normally distributed.
- ACF plot: Few significant lags that confirm there are some autocorrelations left in the residuals.
- The Ljung-Box Test: Most of the points are below red dashed line, we have evidence to reject the null hypothesis that the error terms are uncorrelated.

Overall, based on residual analysis above, normality did not hold for SARIMA $(1,1,1)x(1,0,2)_{(12)}$ model. The model is not very successful to be a good fit.

7. Residual analysis of SARIMA (3,1,0)x(1,0,2)$_{(12)}$ model

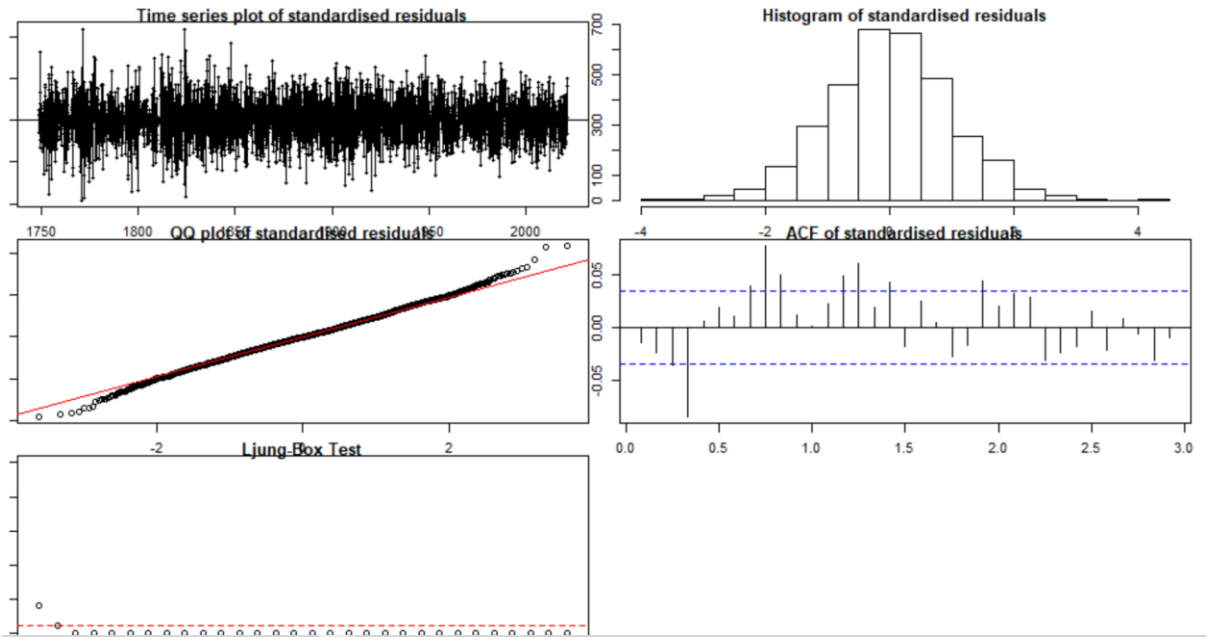


Figure 19 Residual Analysis Plots - FIT 7

```
        Shapiro-Wilk normality test

data:  res.model
W = 0.99752, p-value = 4.11e-05
```

- Time series plot of the residuals shows that it does not capture any trend and hence it becomes a white noise series.
- Histogram of the Residuals showing that the deviation is normally distributed.
- Q-Q plot: We apply quantile-quantile plots for the analysis of normality of residuals. Because majority of points including the extreme ones closely follow the straight line, we cannot reject normality of the error terms in this model.
- Shapiro-Wilk test: With P-value is less than 0.01, so we rejected the null hypothesis that the stochastic component of this model is normally distributed.
- ACF plot: Few significant lags that confirm there are some autocorrelations left in the residuals.
- The Ljung-Box Test: Most of the points are below red dashed line, we have evidence to reject the null hypothesis that the error terms are uncorrelated.

Overall, based on residual analysis above, normality did not hold for SARIMA (3,1,0)x(1,0,2)$_{(12)}$ model**.** The model is also not a good fit.

BEST MODEL**:**

From the overall parameter estimation and residual analysis, we can conclude that the best model that could be fit to our time series data is **SARIMA (3,1,2)x(1,0,2)$_{(12)}$.**

Now let us have a look at how well our model has fit to our time series data,

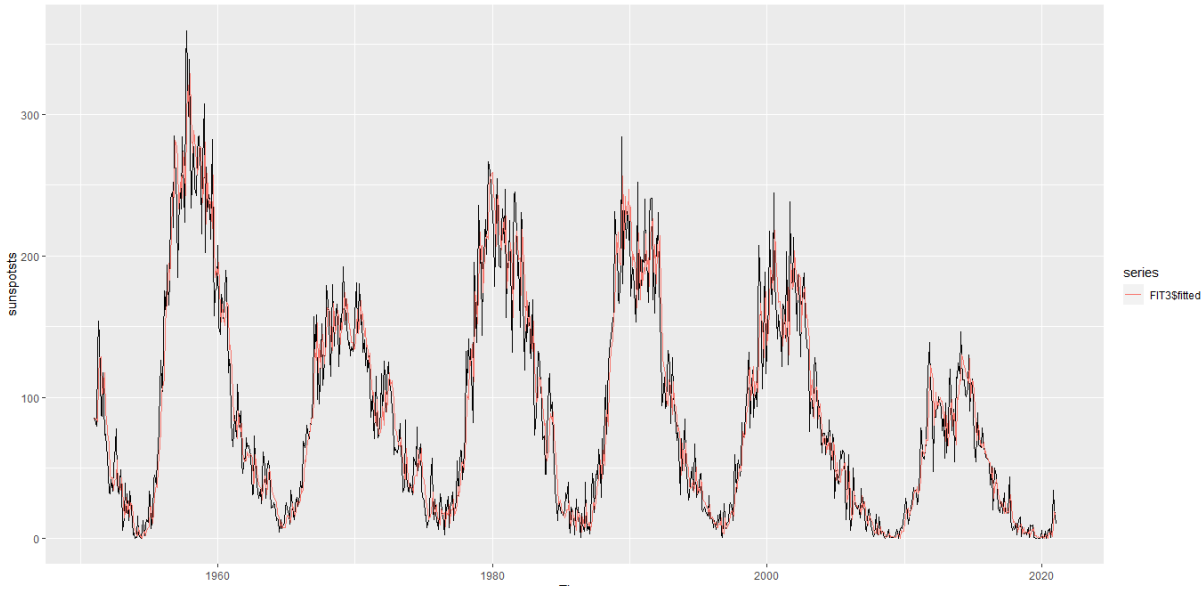


Figure 20 GOODNESS OF FIT

PREDICTION AND FORECAST:

After knowing the best model, we can proceed with prediction and forecast for next years.

Predictions:

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Feb 2021	13.40435	4.608426e+00	27.19976	1.873179e+00	36.60405
Mar 2021	16.46617	5.256350e+00	34.47907	1.926982e+00	46.87918
Apr 2021	18.21906	5.740218e+00	38.35562	2.064448e+00	52.24080
May 2021	18.59077	5.474429e+00	40.21249	1.774954e+00	55.24648
Jun 2021	21.79078	6.595953e+00	46.62378	2.233119e+00	63.83307
Jul 2021	23.54858	6.850058e+00	51.18005	2.177697e+00	70.42080
Aug 2021	24.19763	6.467477e+00	54.30139	1.778905e+00	75.46853
Sep 2021	24.34391	5.821928e+00	56.82961	1.289317e+00	79.94357
Oct 2021	27.60167	6.601211e+00	64.43409	1.461980e+00	90.64073
Nov 2021	29.75629	6.798502e+00	70.54500	1.362913e+00	99.69993
Dec 2021	30.30256	6.237082e+00	74.29892	9.679060e-01	106.05945
Jan 2022	29.87385	5.294089e+00	76.59122	5.182964e-01	110.75530
Feb 2022	32.49553	5.495027e+00	84.41664	4.512886e-01	122.52984
Mar 2022	32.98537	4.899290e+00	88.70286	2.191725e-01	130.00679
Apr 2022	32.97988	4.185897e+00	92.16878	5.695537e-02	136.52251
May 2022	34.27696	3.938580e+00	97.98369	1.206642e-02	146.02313
Jun 2022	35.47952	3.662057e+00	103.77827	-1.964309e-04	155.60687
Jul 2022	36.71361	3.399659e+00	109.76171	-2.043984e-02	165.52589
Aug 2022	36.55784	2.780569e+00	113.34738	-1.423197e-01	172.54622
Sep 2022	37.88532	2.576644e+00	119.71476	-2.648045e-01	183.11964
Oct 2022	38.08981	2.129450e+00	124.10126	-5.382422e-01	191.28832
Nov 2022	38.04767	1.682565e+00	128.07158	-9.465785e-01	198.99487
Dec 2022	39.11376	1.484148e+00	134.21198	-1.283951e+00	209.50629
Jan 2023	39.57067	1.202231e+00	139.25263	-1.789155e+00	218.68965
Feb 2023	41.19388	1.144419e+00	146.32943	-2.075728e+00	230.31293

Figure 21 Predictions of next two years - 2021-2023

- The above plot shows the predictions of the average number of sunspots from the year 2021 to the year 2023.
- The predicted values show us the estimate along with the 80% and 95% confidence intervals.
- For example, for the month Feb 2021, the approximate estimation of the average number of sunspots is 13.4 with 80% confidence interval between 4.6 to 27.2. Also, with 95% confidence interval ranging between 1.87 to 37.6.
- Due to space constraint, only two years' prediction have been reported.

Forecast

However, we can see the forecast for the next 10 years below,

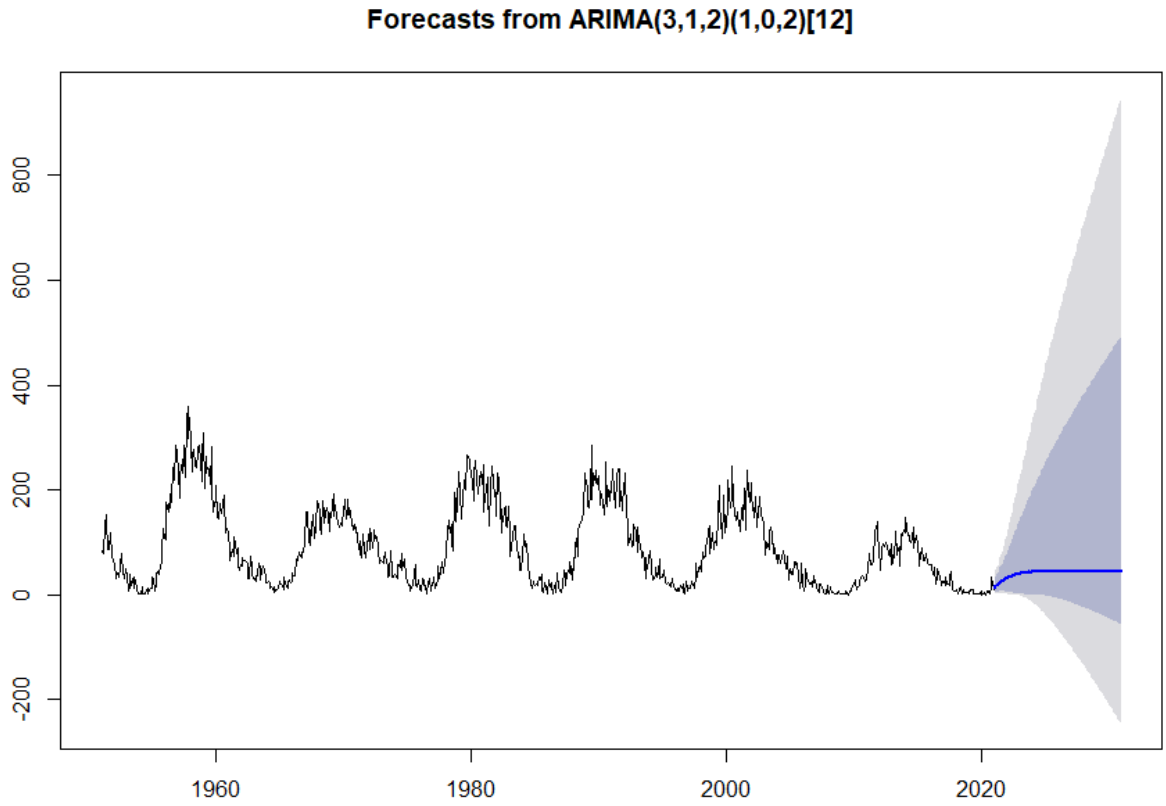


Figure 22 FORECAST OF NEXT TEN YEARS

The above plot shows the forecast of how the mean number of sunspots would vary along the years 2021 to 2031.

Conclusion:

The sunspots timeseries data was found to be non-stationary with a repetitive mid-trend pattern along with MA and AR behavior. There were seasonal and non-seasonal trends observed in the sunspots time series. Due to this, first differencing was applied, and no pattern was visible at seasonal lags. Though there was a presence of seasonality in the time series plot, both ARIMA and SARIMA models were applied on it. Also, since there were no significant seasonal lags observed in the residual analysis of the model, the seasonal

parameters were fixed at (1,0,2) and with the help EACF and BIC table, non-seasonal parameters were found to be (3,1,2) and hence the model ARIMA $(3,1,2)x(1,0,2)_{(12)}$ was obtained. Moreover, predictions were made using this model for next ten years.

REFERENCES:

[1] Sunspots. (2021). Retrieved 13 June 2021, from https://www.kaggle.com/robervalt/sunspots.

APPENDIX A

```
library(TSA)
library(tseries)
library(readr)
library(lmtest)
library(dplyr)
library(lubridate)
library(forecast)

setwd("F:/Subbu/RMIT/sem 3/Time Series Analysis/assign 2")
assign_ds <- read_csv("sunspots.csv",col_names = TRUE)

#FILTER
assign_ds <- tail(assign_ds,841)
assign_ds

# Monthly Mean Total Sunspot Number

sunspotsts <- ts(assign_ds$`Monthly Mean Total Sunspot Number` ,start=c(1951, 1),
end=c(2021, 1), frequency = 12)

plot(sunspotsts, ylab = "Sunspots",  type = "o", main = "Time series plot for the Mean
number of Sunspots along the years")

# AUTO-CORRELATION

plot(y=sunspotsts,x=zlag(sunspotsts),xlab = "Previous Year Number of sunspots",
    ylab= "Average number of sunspots", main = "Scatter plot of Average number of sunspots
vs the first lag ")

#ACF AND PACF
par(mfrow = c(1,2))
acf(sunspotsts,lag.max = 100)
pacf(sunspotsts,lag.max = 100)
par(mfrow = c(1,1))

#RESIDUALS APPROACH
# SARIMA(0,0,0)x(0,1,0)(12)

FIT1=Arima(sunspotsts, lambda = 0.47, order = c(0,1,0), seasonal = c(0,0,0))
coeftest(FIT1)
summary(FIT1)
res1 = rstandard(FIT1)
par(mfrow = c(1,1))
plot(res1)
par(mfrow = c(1,2))
```

```
acf(res1, lag.max = 60)
pacf(res1, lag.max = 60)
par(mfrow = c(1,1))


# SARIMA(1,0,0)x(3,1,2)(12)

FIT2=Arima(sunspotsts, lambda = 0.47, order = c(3,1,2), seasonal = c(1,0,0))
coeftest(FIT2)
res2 = rstandard(FIT2)
plot(res2)
par(mfrow = c(1,2))
acf(res2, lag.max = 60)
pacf(res2, lag.max = 60)
par(mfrow = c(1,1))

# SARIMA(1,0,2)x(3,1,2)(12)
FIT3=Arima(sunspotsts,lambda = 0.47,order = c(3,1,2), seasonal = c(1,0,2)) #SKIPPED(1,0,1)
res3 = rstandard(FIT3)
coeftest(FIT3)
plot(res3)
par(mfrow = c(1,2))
acf(res3,lag.max = 60)
pacf(res3,lag.max = 60)
par(mfrow = c(1,1))

# EACF

eacf(res3)

#p,d,q - (0,1,0), (0,1,1), (1,1,1)

# SARIMA(1,0,2)x(0,1,0)(12)
FIT4=Arima(sunspotsts,lambda = 0.47,order = c(0,1,0), seasonal = c(1,0,2))
res4 = rstandard(FIT4)
plot(res4)
par(mfrow = c(1,2))
acf(res4,lag.max = 60)
pacf(res4,lag.max = 60)
par(mfrow = c(1,1))


# SARIMA(1,0,2)x(0,1,1)(12)
FIT5=Arima(sunspotsts,lambda = 0.47,order = c(0,1,1), seasonal = c(1,0,2))
res5 = rstandard(FIT5)
plot(res5)
par(mfrow = c(1,2))
```

```
acf(res5,lag.max = 60)
pacf(res5,lag.max = 60)
par(mfrow = c(1,1))


# SARIMA(1,0,2)x(1,1,1)(12)
FIT6=Arima(sunspotsts,lambda = 0.47,order = c(1,1,1), seasonal = c(1,0,2))
res6 = rstandard(FIT6)
plot(res6)
par(mfrow = c(1,2))
acf(res6,lag.max = 60)
pacf(res6,lag.max = 60)
par(mfrow = c(1,1))

#BIC
RES_bic = armasubsets(y=res3 , nar=5 , nma=5, y.name='Average number of sunspots',
ar.method='ols')
plot(RES_bic)

# SARIMA(1,0,2)x(3,1,0)(12)

FIT7=Arima(sunspotsts,lambda = 0.47,order = c(3,1,0), seasonal = c(1,0,2))
res7 = rstandard(FIT7)
plot(res7)
par(mfrow = c(1,2))
acf(res7,lag.max = 60)
pacf(res7,lag.max = 60)
par(mfrow = c(1,1))

# RESIDUAL ANALYSIS FUNCTION

residual.analysis <- function(model, std = TRUE,start = 2, class =
c("ARIMA","GARCH","ARMA-GARCH", "fGARCH")[1]){
  library(TSA)
  library(FitAR)
  if (class == "ARIMA"){
    if (std == TRUE){
      res.model = rstandard(model)
    }else{
      res.model = residuals(model)
    }
  }else if (class == "GARCH"){
    res.model = model$residuals[start:model$n.used]
  }else if (class == "ARMA-GARCH"){
    res.model = model@fit$residuals
  }else if (class == "fGARCH"){
    res.model = model@residuals
```

```
  }else {
    stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")
  }
  par(mfrow=c(3,2))
  plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of
standardised residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
  qqline(res.model, col = 2)
  acf(res.model,main="ACF of standardised residuals")
  print(shapiro.test(res.model))
  k=0
  LBQPlot(res.model, lag.max = 30, StartLag = k + 1, k = 0, SquaredQ = FALSE)
  par(mfrow=c(1,1))
}

# RESIDUAL ANALYSIS

par(mar = c(1,1,1,1))

residual.analysis(model = FIT1)
residual.analysis(model = FIT2)
residual.analysis(model = FIT3) #BEST FIT - SARIMA(1,0,2)X(3,1,2)(12)
residual.analysis(model = FIT4)
residual.analysis(model = FIT5)
residual.analysis(model = FIT6)
residual.analysis(model = FIT7)

par(mar = c(1,1,1,1))

# HOW WELL THE MODEL FITS

autoplot(sunspotsts) + autolayer(FIT3$fitted)

# FORECAST

Ffrc = forecast(FIT3,h=120)
Ffrc
plot(Ffrc)
```