TIME SERIES ANALYSIS REPORT

Abstract

This report is aimed to analyze time series data by using different statistical methods for model specification and investigate the farecasting capacity of each model. Autoregressive moving average model (ARMA), autoregressive integrated moving average model (ARIMA) and the seasonal autoregressive integrated moving average model (SARIMA) were used to forecast the observations. The following study is in relation to the question being analyzed, how to apply above mentioned statistical models to forecast the time series? Hence, three different time series data sets were used in this study to explain the above study question. Time series data set was considered in three different topics in three different frequencies such as annual, quarterly and monthly to provide a diverse background to the study. The data was analyzed using the R-programming language and results was interpreted accordingly. The study concludes with limitations and directions for future analysis.

Introduction

Much of the attention was given for analysing time series data sets by using different statistical models to determine the forecasting ability of the historical data. Hence, this study aims to fill the knowledge in proper application of those statistical models for time series in different topics and different frequencies to provide enthusiasm on analysing further.

The following study is in relation to the question being analysed, that is, how to apply above mentioned statistical models to forecast the time series?

This analysis scope is mainly considered on application of autoregressive moving average model (ARMA), autoregressive integrated moving average model (ARIMA) and the seasonal autoregressive integrated moving average model (SARIMA) to the three time series data used in this study.

Statistical Methodology

The data was analysed using the R-programming language and results was interpreted accordingly. Further, autoregressive moving average model (ARMA), autoregressive integrated moving average model (ARIMA) and the seasonal autoregressive integrated moving average model (SARIMA were used to interpret the findings.

As such, different hypothesis were tested and different statistical tests were performed as appropriately to determine stationary, residuals analysis, model fitting and forecasting.

Overview of the Data Sets and Data Processing

Three data sets have been used for this analysis and those are explained below.

Application of SARIMA model to a quarterly data set is describe in the first section. The first data set used in this assignment is called `OS visits to UK: All Visits Thousands-NSA', which is about the number of overseas visits to UK during the period of 1980 to 2019. The quarterly data available in this data set has been used that includes 684 observations and 2 variables. Further, missing values have not been examined in the data set. One outlier was identified and removed from the data set which is relevant to the data of first quarter (Qtr) of year 2020, which is possibly due to travel restrictions due to Covid 19. Transformation mechanisms have not been used since the above time series doesn't indicate abnormal variations and pattern. Hence, Boxcox and log transformations were not used.

Summary statistics for the data set indicates, the average number of Overseas visits to UK is 6,270 ('Thousands), while minimum is 1,920 ('Thousands) and maximum is 11,899 ('Thousands).

Application of ARIMA model to annual data set is describe in the second section. The second data set used in this assignment is called `Gross Domestic Product at market prices: CP: NSA £m', which is about the number of GDP market value for the period of 1948 to 2021. The annual data available in this data set has been used that includes 342 observations and 2 variables. Missing values have not been examined in the data set. Outliers also not examine in this data set as there is a gradual increase over the time period. The above time series consist with monetary data which is GBP in millions. It is apparent that the monetary value is reduced over time. Hence, GDP per-capita, adjusting the data for inflation using CPI are possible transformation options for this data set to smooth the effect of monitory value. However, those transformations were not applied at this point to the time series in this analysis.

Summary statistics for the second data set indicates, the average of GDP value is 696,582 (£m) for the period given, while minimum is 11,425 ((£m)) and maximum is 2,317,667 (£m).

Application of SARIMA model to a monthly data set is describe in the third section. The third data set used in this assignment is called `UK: Women Employment Fulltime: Aged 16 and over (Thousands): NSA', which is basically about the statistics on fulltime employed women in UK during the period of 1993 to 2020. The monthly data available in this data set has been used that includes 347 observations and 2 variables. Missing values and outliers have not been examined in the data set. Transformation mechanisms have not been used since the above time series doesn't indicate abnormal variations and pattern. Hence, Boxcox and log transformations were not used.

Summary statistics for the third data set indicates, the average employment count is 7637 ('Thousands), while minimum is 6343 ('Thousands) and maximum is 9590 ('Thousands).

Source:
(https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/timeseries/i46h/lms)

(https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/bktl/pn2)

(https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/timeseries/gmaa/ott)

Application of SARIMA to a quarterly data set

Exploratory Data Analysis

Original time series can be express as a function that explain the trend, seasonality effect and a stationary component(random component) as white noise. Hence, it was performed an analysis below to identify the trend and seasonality components in the created time series object simply by plotting the original time series data as below Figure 1.

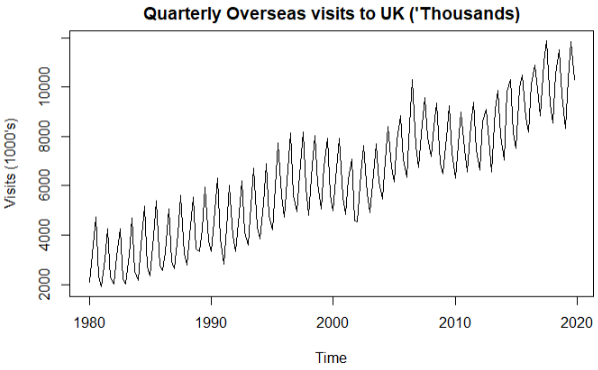
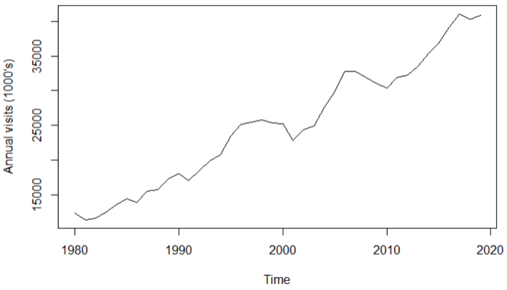


Figure 1 : Time series of `OS visits to UK'

Above Time plot illustrate followings; that there is a increasing trend (upward trend) as the time goes up, the number of visits also growing up in general. The time series displays a very regular pattern called seasonality. Further, seasonality component indicate a strong seasonal pattern between the quarters.This has been analyse in detailed at the plots relevant to seasonality effect.

The variance (size of the visits) is not increasing as the time (level of the series) increases. Hence, Box-Cox transformation has not been applied for this time series. All the peaks and drops are not in same size, as there is some viability. As such, it can assume that the Time series variance is not constant over the time. Plotting time series object enables to identify the trend and seasonality component of the time series easy. Hence following plots have been generated.



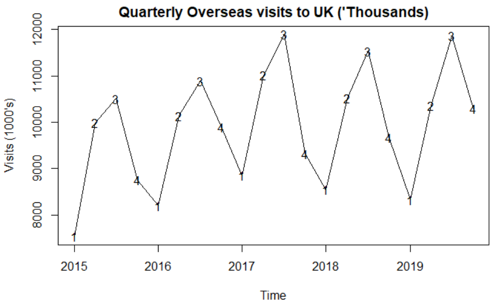


Figure 2: Aggregated data over the years

Figure 3: Plotting a portion of time series object

The aggregate (annual) data was plotted (Figure 2) to see the trend pattern of annual data and which indicates a upward trend (slighly closer to a deterministic trend) as the time goes up the visits also growing up. Time series object has been plotted for a portion of recent time slots (Figure 3) to visibly see the pattern of the time series more clearly between quarters. Above time plot indicates a regularly repeating pattern of high and lows related to quarters of the year. In this case, it could clearly seen a drop in first quarter (1), then a gradual increase upto third quarter (3). The graph also shows that there is a drop again in fourth quarter (4).As such, it provides a sense of seasonality component associated with this time series.

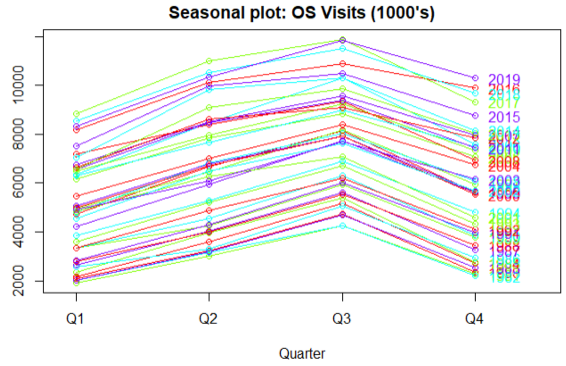


Figure 4 : Plot of seasonality

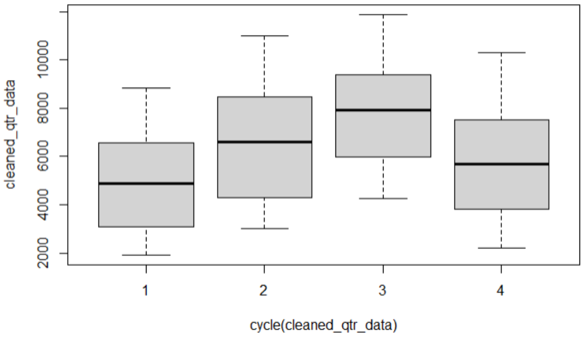


Figure 5 : Boxplot for four quarter

Above seasonal plot in figure 4, indicate the underlying seasonal pattern more clearly. The data from each season are overlapped for four quarters. There is a drop in first quarter (Qtr), then a gradual increase upto third Qtr and again a drop in fourth Qtr. The seasonal pattern between quarters as an increase in summer and then decrease in winter for the number of visits.

The Boxplot in Figure 5, is a form of plot enables the underlying seasonal pattern to be seen clearly, and also shows the changes in seasonality over time. There is a pattern when observing the median line of the each Boxplots which at the beginning of the year tend to decrease. Then at the middle of the year which correspond to summer time there is a increase and again move down at the end of the year which is winter. Highest median value and the range is for the third frequency (Qtr 3) where as the lowest is for the first frequency (Qtr 1).

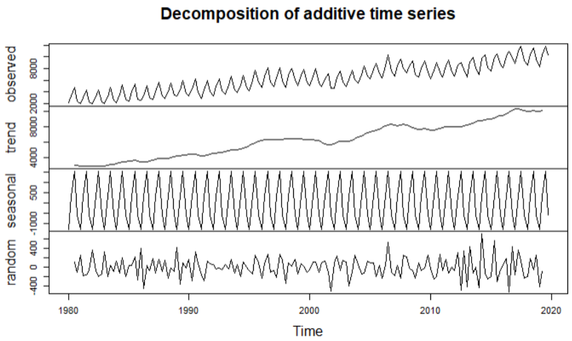


Figure 6 : Decomposition of time series

Additive model was used as the time series since the variance (size of the visits) is not increasing as the time (level of the series) increases by (variance volatility is not increasing with the time).Decomposition plot in Figure 6, illustrate the trend, seasonal and random component of the time series separately in a graphical manner as above. The underlying time series is a non-stationary time series as it has a trend or a seasonal component and requires differencing to transform it to a stationary time series.

Model fitting and Forecasting

Stationary time series is devoid of trend, seasonal patterns and white noise. A time series is said to be stationary if it holds the following conditions true.

- The expected value (mean value) of time-series is constant over time (which implies, the trend component is nullified)
- variance is constant which oscillate around a constant value, and the correlation depends on the time lag, but not the absolute.

It is important to note that the process defined in this way is weakly stationary. ACF plot and hypothesis testing (ADF & KPSS) was performed to identify that the time series of Overseas visits to UK is stationary or not.

1.9.1 Checking for stationary - Autocorrelation function (ACF)

The autocorrelation coefficients are plotted to show the autocorrelation function or ACF. ACF plots observe if the data has a trend and a seasonal component. The plot is also known as a correlogram which is in below.

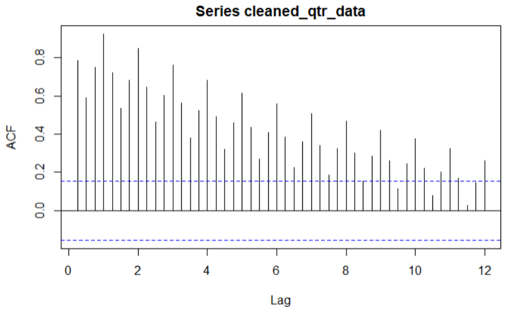


Figure 7 : Autocorrelation function for original time series

Above plot of ACF indicates significant positive correlations. The autocorrelations decreases slowly (irrespective of seasonal pattern) as the number of lags increases. Further, there is a clear seasonal pattern is visible as there are some peaks at every four lags. The autocorrelations are gradually decreasing after the every peak as the lags increases. A combination of these effects could be seen, when data are both trended and seasonal.

Hence, ACF indicates trended time series with a seasonality effect. In a stationary time series, the ACF will drop to zero relatively quickly as it doesn't depends on time, while the ACF of non-stationary data decreases slowly as in the above plot. Further, more than 5% of spikes are outside the bounds. As such, the above ACF indicates a that the attributes of non-stationary time series or not a white noise.

Further, it is also useful to quantify the evidence of non-stationarity via hypothesis testing. Hence, the Dickey -Fuller test and KPSS tests has been used in below. The p-value of above ADF test is 0.1973, which is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that there the time series is not stationary. The p-value of the KPSS test is 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected and conclude that the time series is not a stationary. Hence, stationary through differencing is explain below.

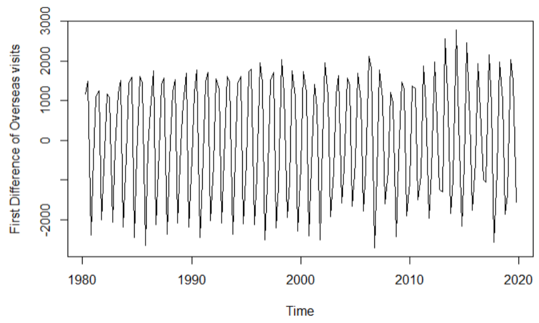


Figure 8 : Time plot for First difference

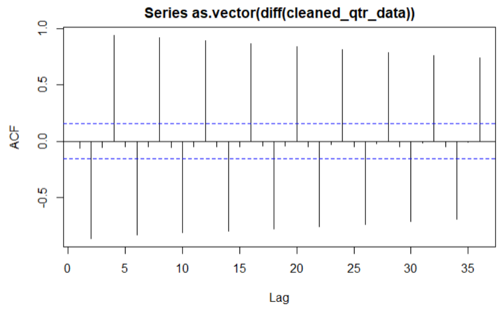


Figure 9 : ACF for First difference

A time plot after the first difference is illustrate above in figure 8. In this case, it is clear that the trend component has been eliminated from the series and values oscillate around a constant level. However, there is still regular pattern is visible which repeat every year. That is an indication of the seasonal pattern which require seasonal differencing to convert the time series to a white noise.

Further, above ACF plot (Figure 9) indicates autocorrelation function for the first Difference. There are many significant spikes with a pattern. Hence, second difference is applied to the time series to remove the seasonal effect.

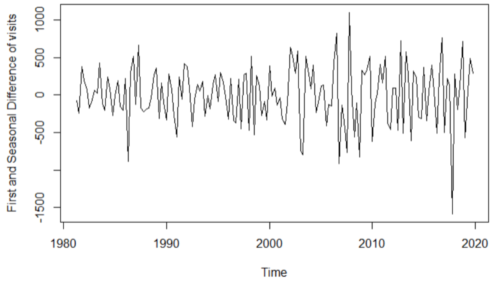


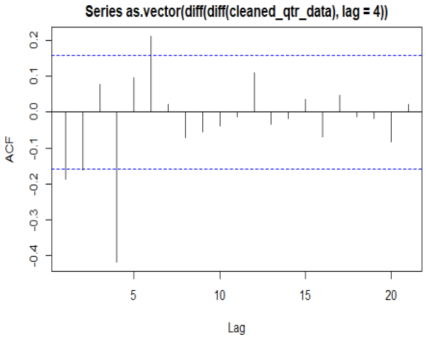
Figure 10 : First and Seasonal Difference

The above Figure 10, displays the time plot after application of both first difference and seasonal difference. As such, there is no trend component, since the values are oscillated around a constant level. Further, the seasonal pattern has been eliminated and a random pattern is visible. Accordingly, both the trend and seasonal component has been eliminated from the time series. Further, the evidence of non-stationarity could be quantified via hypothesis testing. Thus, Dickey -Fuller test has been applied below.

Checking for stationary by using Augmented Dickey-Fuller unit-root test (ADF) test which ADF test is 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is a stationary. The KPSS test in which is 0.1, which is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that there the time series is a stationary.

Afterwards, the following ACF, PACF and EACF plots have been analyzed to determine the parameters initially.

Model Specification



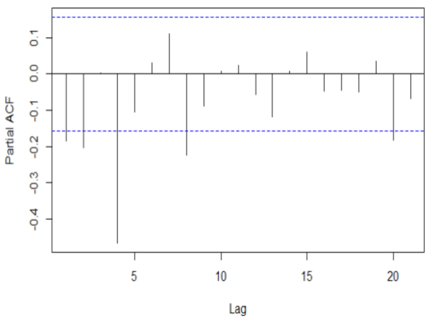


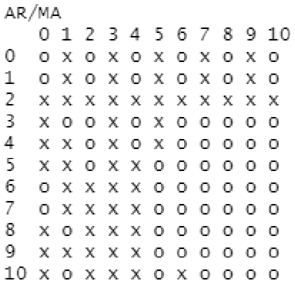
Figure 11 :ACF

Figure 12 :PACF

Above ACF function (Figure 11) illustrate that approximately 95% of the spikes are within the interval of confidence. However, there are very few significant autocorrelations are out from the interval of confidence. The spikes at lag 1,4, and 6 are significantly different from zero and rest of all lags are zero or closer to zero. Hence, it could assume that the time series is a is a stationary. Above ACF plot has been used to determine moving average (MA) component of the model specification. As such, ACF illustrate both a non-seasonal moving average component (q) and seasonal moving average component (Q). Hence, it could assume 'q' as 1 or 2, and 'Q' as 1. The first difference (d) was set as 1.

PACF function indicates in the Figure 11, has very few significant autocorrelations are out from the interval of confidence. However, it can assume that the 95% is within the interval of confidence. The spikes at lag 1,2,4,8,and 20 are significantly different from zero and rest of all lags are zero or closer to zero. Hence, it could assume that the time series is a is a stationary. PACF plot illustrate autoregressive (AR) component of the of model specification. The time series in this analysis consist with both non-seasonal autoregressive component (p) and seasonal autoregressive component (Q). Significant Peaks tend to be four lags apart. Hence, by examining the above PACF plot,it could assume 'p' as 1 or 2, and 'P' as 1 or 2. (Lag 20 has not considered as not all the 20 parameters are significant and it will lead to increase the errors in the model). The seasonal difference (D) was set as 1.

By examining the EACF plot(Figure 13) to determine a appropriate model is complicated as it does not have a clear visible pattern.



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

Figure 13: EACF

Parameter Estimation

As such, Seasonal ARIMA(p,d,q)×(P,D,Q)s model would be appropriate to apply for the time series in this analysis where nonseasonal orders (p, d, q),seasonal orders (P, D, and Q) and seasonal period s. Thus, ACF, PACF plots and AIC, BIC criteria has been used in the process of model specification & parameter estimation of the model. Hence, list of possible models more than 10 was tested against the AIC criteria and finally selected two models. As per the above output results, the lowest AIC value which is 2,229.78 for "pq0201" provides and then the second lowest AIC value of 2,229.82 for "pq1111". Hence, it can assume 'pq0202' as the best model based on the AIC value and the second best as 'pq111'. The attributes of the model are described below.

The lowest value of 2,229.78 is given by 'pp0201' which has 3 parameters and the coefficients of parameters (MA1, MA2 and SMA1) is -0.3031, -0.1229 and -0.7351 respectively. Square error also quite small for all parameters (0.0819, 0.0780 and 0.0728 respectively). The second lowest value of 2,229.82 is given by 'pp1111' which has 4 parameters and the coefficients of parameters (MA1,MA2,SMA1 and SMA1) is 0.3350, -0.6629, 0.1792 and -0.8181 respectively. However, square error also quite small for all parameters but slightly above to the previous model.

Further, the number of parameters are lower in the model 'pp0201' and lesser the parameters is better. By comparing two models, it could assume that the pp0201 is the best model in terms of AIC value and number of parameters.As such, two models (pp0201 and pp1111) could be specify as below.

- Best model out of the evaluated list of models: **Seasonal ARIMA(0,1,2)×(0,1,1)4**
- Second best model out of the evaluated list of models **: Seasonal ARIMA(1,1,1)×(1,1,1)4**

Further, BIC criteria have used to evaluate the models which has the lowest AIC values. As such, BIC criteria provide lowest value for "pq0201" which is 2,243.953. Hence, both tests suggest the "pq0201" appears to be an more appropriate model for the time series that is seasonal ARIMA (0,1,2)×(0,1,1)4

Residuals should be a white noise which is a weaker condition of stationary. Accordingly analysis is performed for residuals as below for the above selected two models; seasonal ARIMA (0,1,2) ×(0,1,1)4 and seasonal ARIMA(1,1,1)×(1,1,1)4

Residual Analysis (Estimate of White noise)

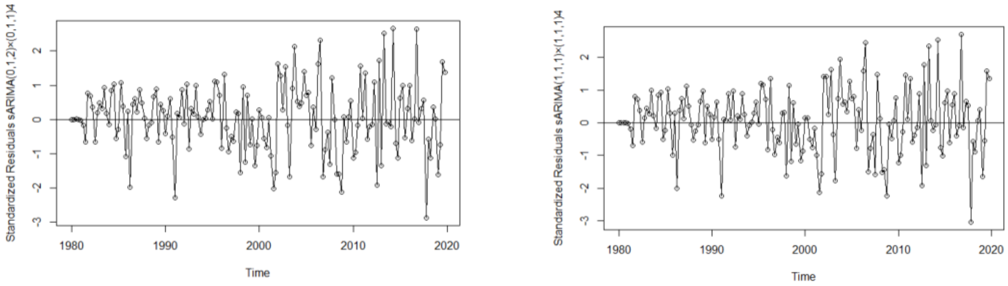


Figure 14 : Reseduals plots for 2 selected models

The above residual plots in Figure 14, illustrate that the majority of residuals are oscillate around zero and approximately 95% of data fall between +2 & -2. As residuals should be a stationary which is independent and identically distributed. Hence, ACF of residuals should be close to zero and 95% of lags should be inside the interval of confidence.

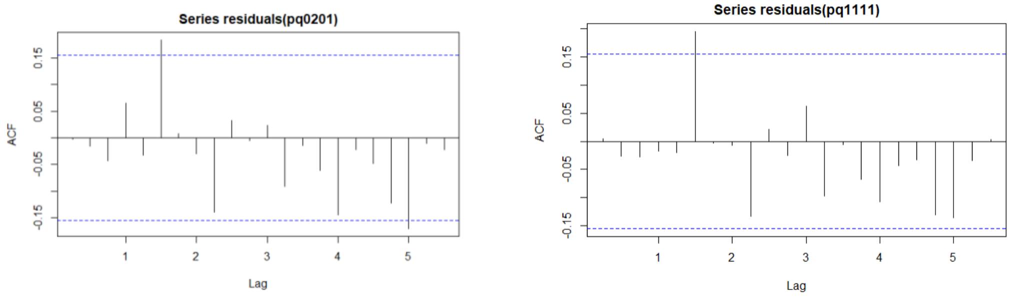


Figure 15: Resedual plots for two models

ACF function for residuals indicates that only one or two autocorrelation are out from the interval of confidence. Hence, that the 95% is within the interval of confidence and are approximately zero or closer to zero. As such, it appears to be no significant autocorrelation in the residuals and the residuals are uncorrelated. P-value of above Box-Pierce test and the Lung-Box test are not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest for Residuals or error terms are uncorrelated.

The p-value of both ADF test are below 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is a stationary. This confirms the residuals are stationary or white noise.

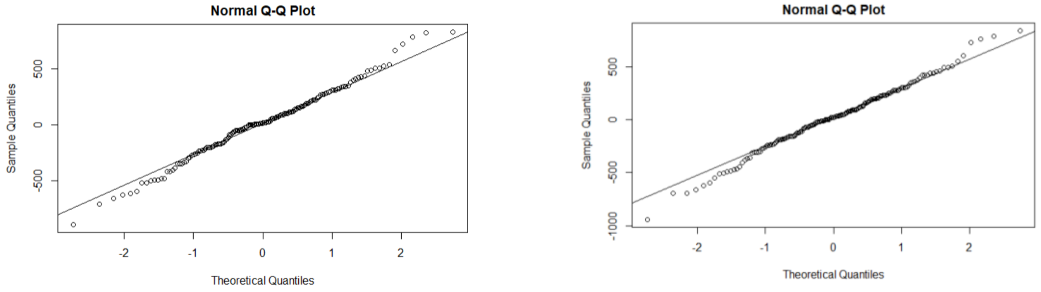


Figure 16: Normality of the Residuals - QQ plots

Above plots in Figure 16 illustrate that the quantiles of the residuals is close to the normal distribution in general. However, few deviations are noticeable at the edge of the line. Hence, A statistical test has been performed to check the normality of the distribution.

Shapiro-Wilk test of normality has a test statistic p-value is 0.417 and HO should not be rejected. Hence, the residuals are normally distributed. Residuals analysis illustrate that the residuals are stationary/white noise (weaker conditions), uncorrelated (correlation=0, independent) and normally distributed for both the models. Hence, the summary plot for residual analysis is illustrate in Figure 17.

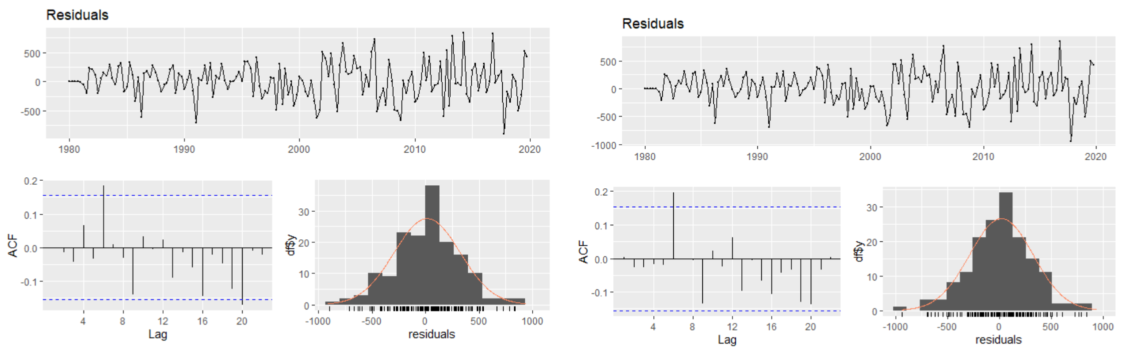
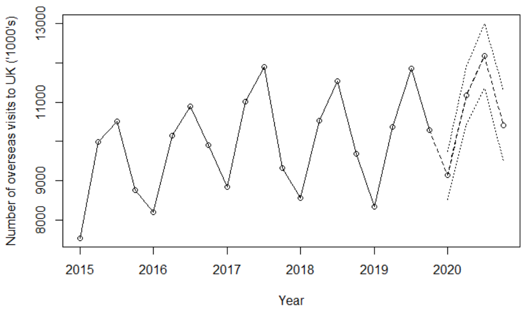


Figure 17 : Resedual Plots

Over fitting exercise was carried out by adding more petameters to evaluate whether the selected model is overfitting. As such, one parameter has added to our model as below.New model has four parameters and the estimation for extra parameter is -0.008 which is very close to zero. Hence, adding a new parameter (MA3) is insignificant and made the model ineffective. Further, the AIC value for the new model (pq0301) is 2,231,77 which is slightly higher than the original model 'pq0201'. Hence, It provide a sign that the selected model is a overfitted model.

Model Forecasting

Forecasting for the model pp0201 which is (seasonal ARIMA (0,1,2)×(0,1,1)4) is as below.



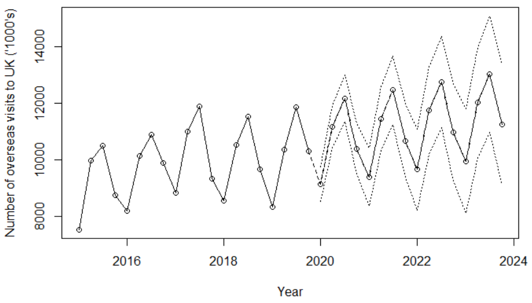
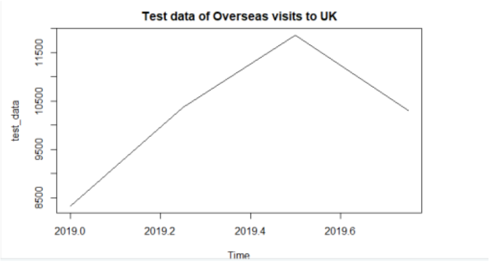


Figure 18: Forecasting for 1 year

Figure 19: Forecasting for 4 years

Forecasting was done for next 4 and 16 quarters respectively. It could examine that the forecasted values of the model (seasonal ARIMA(0,1,2)×(0,1,1)4) is similar to the pattern of the previous quarters of the year. Hence, model was fitted to forecast the values.The time plot for next 16 quarters indicate that the interval of confidence increases as the period increases which leads predictions become less precise. Hence, forecasting for quite lengthy periods tend to have increased interval of confidence.

A forecast "error" is the difference between an observed value and its forecast. Accuracy of the model prediction depends on the length of the trained data. Hence, the data set used for this analysis has quite lengthy period of past data starting from Qtr1 of 1980 to Qtr 4 of 2019. In general, above two test results for forecast errors provide less values for the pq1111 model which is seasonal ARIMA(1,1,1)×(1,1,1)4. Hence, based on the analysis of forecast errors, it assume seasonal ARIMA(1,1,1)×(1,1,1)4 as the best model out of the tested models.



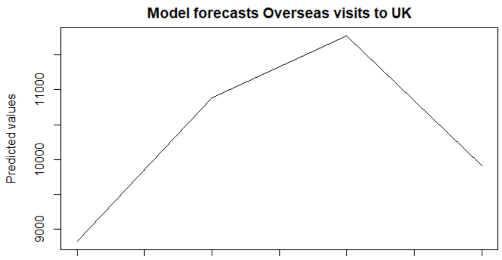
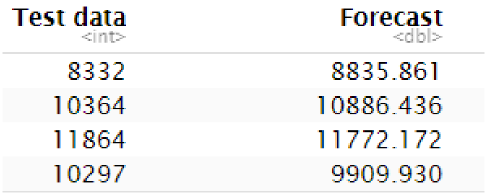


Figure 20 : Test data set

Figure 21: Model predictions for test data

Above, output illustrate the values for the test data and model predicted value for test data along with the forecasted values for each quarter of year 2020. As such, the predicted values of the fitted model is closer to the actual values. Further, below plots illustrate the patterns relevant to the above data. Thus, the model predictions for existing data is indicate in the Figure 22 below..



Test data	Forecast
<int>	<dbl>
8332	8835.861
10364	10886.436
11864	11772.172
10297	9909.930

Figure 22 : Comparison of Predictions for the year 2020

Conclusion:

Time series for number of overseas visits to UK was analysed in this study. As such, time series was convert to a stationary time series, develop list of possible models, model fitting and , residual analysis was performed. Finally, Seasonal ARIMA(1,1,1)×(1,1,1)4 was fitted to forecast the number of overseas visits to UK. Then, apply that model to train and test data set of the same time series and evaluate the prediction criteria. As such, the model follows the similar pattern to that of past data and predict values closer to the test value set.

This model can be improved by application of some machine learning prediction algorithms to predict the values almost similar er to the test values. Further, this data set is related to the Leisure and Tourism industry, which was affected drastically during recent past. Hence, another model could be develop to track the during_Covid and post_Covid values.

Application of ARIMA to an annual data set

Explanatory data analysis

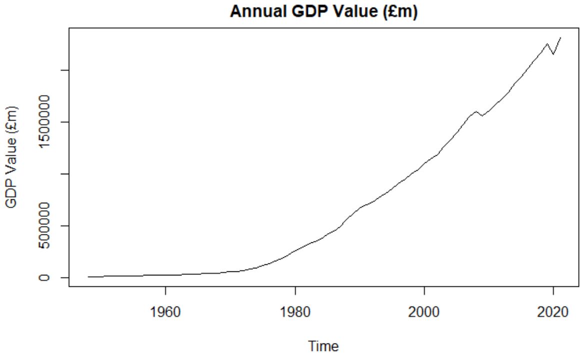


Figure 23 : Plotting the time series object

Above Time plot in Figure 23, illustrate followings. There is a increasing trend (upward trend) as the time goes up, GDP value also growing up in general. It can argue as a quadratic trend also. The time series do not indicate any sense of seasonality. There is a slight drop in the trend pattern in 2009 and 2020. This is mainly due to the economic downturn/ressession in 2009 and Covid 19 effect.

Task 2 – Model fitting and Forecasting

It is important to note that the process defined in this way is weakly stationary. ACF plot and hypothesis testing (ADF & KPSS) was performed to identify that the time series object is stationary or not. The autocorrelation coefficients are plotted to show the autocorrelation function or ACF. ACF plots observe if the data has a trend and a seasonal component. The plot is also known as a correlogram which is in below.

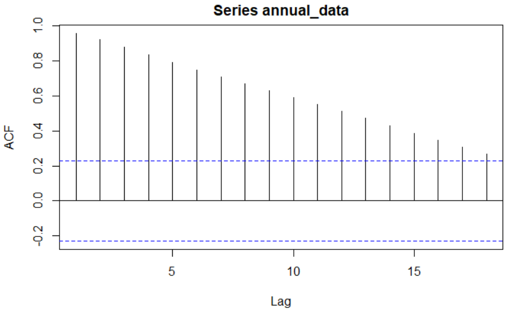
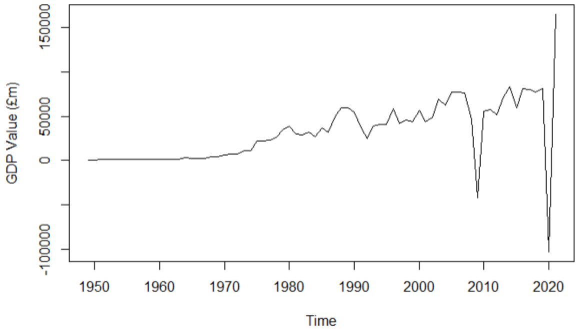


Figure 24: Autocorrelation function for original time series

Above plot of ACF indicates significant positive correlations. The autocorrelations decreases slowly as the number of lags increases. Further, there is no any clear seasonal pattern visibly in the ACF. Hence, ACF indicates trended time series. In a stationary time series, the ACF will drop to zero relatively quickly as it doesn't depends on time, while the ACF of non-stationary data decreases slowly as in the above plot. Further, more than 5% of spikes are outside the boundaries. As such, the above ACF indicates the attributes of non-stationary time series or not a white noise.

Further, it is also useful to quantify the evidence of non-stationary via hypothesis testing. Hence, the Dickey -Fuller test and KPSS tests has been used in below. The p-value of above ADF test is 0.8037 which is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that there the time series is not stationary. Further, The p-value of the KPSS test is smaller than 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is not a stationary. As such, It is apparent that the above time series is a non-stationary since it has a trend component. Hence, it is required to convert to a stationary time series to construct a appropriate model for forecasting.



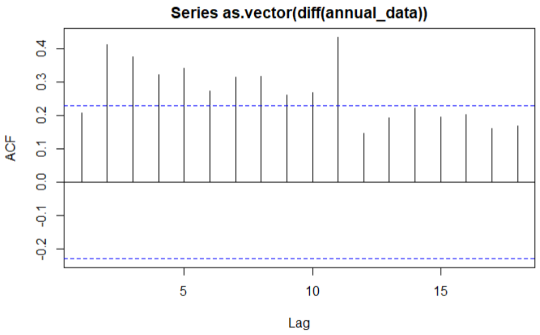


Figure 25: First difference time plot

Figure 26: First difference ACF

A time plot after the first difference is illustrate above. In this case, there indicate some random pattern, however, trend is still visible in the series. That is an indication of the for the requirement of second differencing to convert the time series to a white noise. Above ACF plot indicates autocorrelation function for the first difference. There are many significant spikes beyond the interval of confidence. Hence, second difference is applied to the time series to remove this effect.

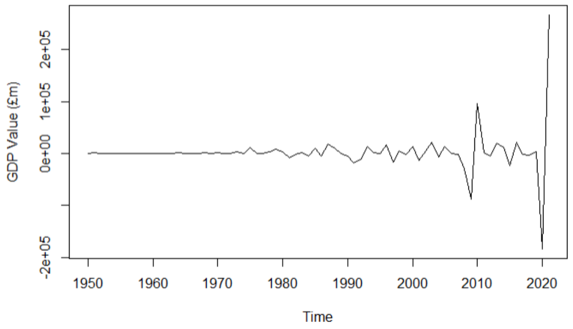


Figure 27: First and Second Difference

The above Figure 27displays the time plot after application of both first difference and second difference. As such, there is no trend component, since the values are oscillate around a constant level. Further, few ups and downs are noticed in the time series which is insignificant. The evidence of non-stationarity could be quantified via hypothesis testing further. Thus, Dickey -Fuller test has been applied below.

The p-value of above ADF test is smaller 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is a stationary. The p-value of above KPSS test is 0.1, which is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that there the time series is a stationary.

Model Specification

The following ACF, PACF and EACF plots have been analysed to determine the parameters initially.

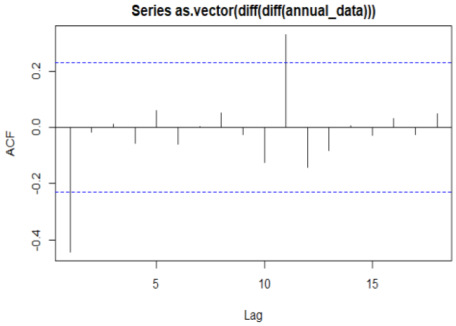


Figure 27 : ACF

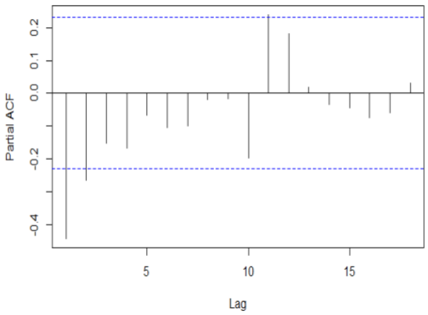
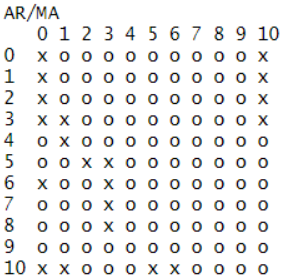


Figure 27 : PACF

Above ACF function in figure 27, illustrate that approximately 95% of the spikes are within the interval of confidence. However, there are very few significant autocorrelations are out from the interval of confidence. That is lag 1 and 11 which is significantly different from zero and rest of all lags are zero or closer to zero. Hence, the time series is a is a stationary.Above ACF plot could be used to determine the moving average (MA) component of model specification. As such, above ACF illustrate non-seasonal moving average component (q) which could be assume as 1. The difference (d) was set as 2 (d=2), since two differencing were applied to the series.

PACF function indicates in Figure 27, that very few significant autocorrelations are out from the interval of confidence. However, it can assume that the 95% are within the interval of confidence. That is lag 1,2 and 11 which is significantly different from zero and rest of all lags are zero or closer to zero. Hence, the time series is a is a stationary. Thus PACF plot illustrate autoregressive component (AR) component of model specification. Hence, by examining the above PACF plot, it could assume non-Seasonal autoregressive component (p) as 1 or 2. The difference (d) was set as 2 (d=2), since two differencing were applied to the series.

The EACF table illustrate a triangular pattern of zeroes with the top-0,1 point. Thus, EACF plot illustrate AR/MA as (0,1) is as an appropriate for the model specification.



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

Figure 28: EACF

Parameter Estimation

Accordingly, ARIMA(p,d,q) model would be appropriate to apply for the time series in this analysis where nonseasonal orders (p, d, q). Thus, ACF, PACF plots and AIC, BIC criteria has been used in the process of model specification & parameter estimation of the model. Accordingly, list of possible models have been evaluated to identify the most appropriate model for the time series.

As per the above output results, the lowest AIC value which is 1,688.78 for "pq11", the second lowest AIC value of 1,690.18 for "pq21" and the third lowest AIC value of 1,692.45 for "pq01". Hence, it can assume 'pq11' as the best model based on the AIC value and the second best as 'pq21'. The three models which has the lowest AIC value was tested for BIC criteria. Accordingly, 'pq11' has the lowest BIC value, then 'pq01' and 'pq21' respectively. Thus, both AIC and BIC criteria suggest 'pq11' as the best model.

The attributes of the model are describe below. The lowest AIC & BIC values given by 'pq11' model which has 2 parameters (ar1 and ma1) and the coefficients of parameters are -0.3516, and -0.7999 respectively. Square error also quite small for the two parameters (0.1440, and 0.0721 respectively). Number of parameters are lower in the model 'pq11' than the 'pq21' model and lesser the parameters is better the model. The second lowest AIC value is for 'pq21' model which has three parameters (ar1, ar2 and ma1) and coefficients of the parameter are -0.3351, 0.1594, -0.8268 respectively. Square error also quite small for the parameters and AIC value as 1690.18.

As such, two models (pq11 and pq01) could be specify as below.

- **Best model out of the evaluated list of models : ARIMA(1,2,1)**
- **Second best model out of the evaluated list of models : ARIMA(2,2,1)**

Residual Analysis (Estimate of White noise)

Residuals should be a white noise which is a weaker condition of stationary. Accordingly analysis is performed for residuals as below for the above selected two models; ARIMA(1,2,1) and ARIMA(2,2,1)

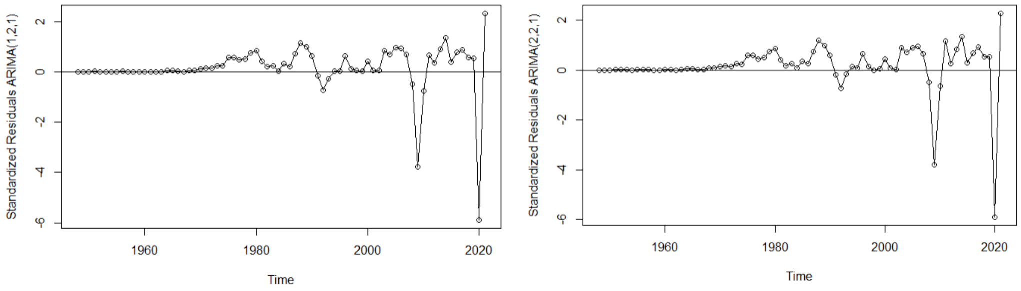


Figure 29 : Reseduals plots for two models

The above residual plots illustrate that the majority of residuals are oscillate around zero and approximately 95% of data fall between +2 & -2. However, the reseduals do not indicate a pattern, but the distributed randomly. As residuals should be a stationary which is independent and identically distributed. Hence, ACF of residuals should be close to zero and 95% of lags should be inside the interval of confidence.

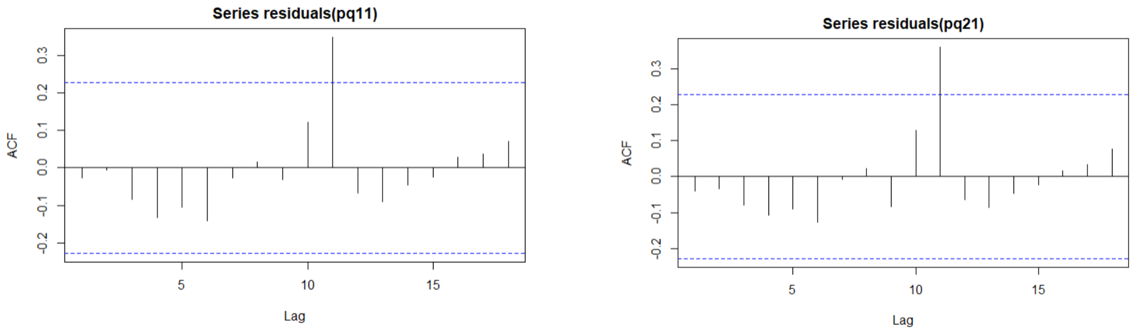


Figure 30 : ACF for two models

ACF function for residuals indicates that only one or two autocorrelation are out from the interval of confidence. Hence, that the 95% is within the interval of confidence and are approximately zero or closer to zero. As such, it appears to be no significant autocorrelation in the residuals and the residuals are uncorrelated. The p-value of above Box-Pierce test and the Ljung-Box tests are not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest for Residuals or error terms are uncorrelated.

The p-value of both ADF test are below 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is a stationary. This confirms the residuals are stationary or white noise.

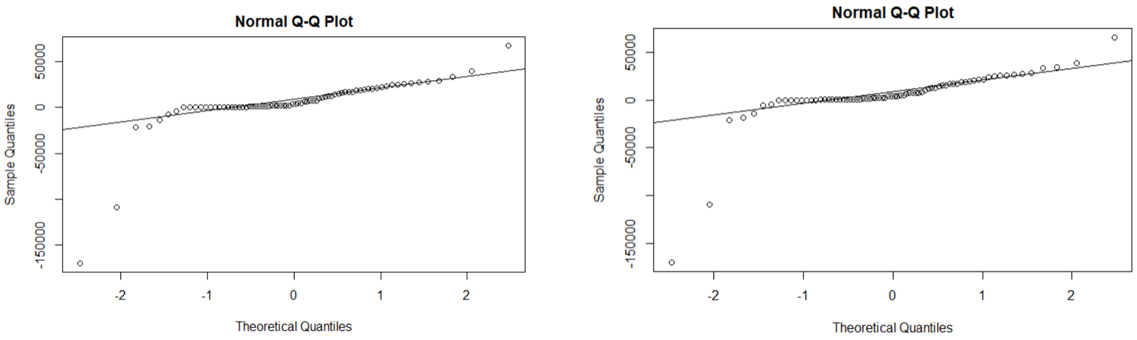


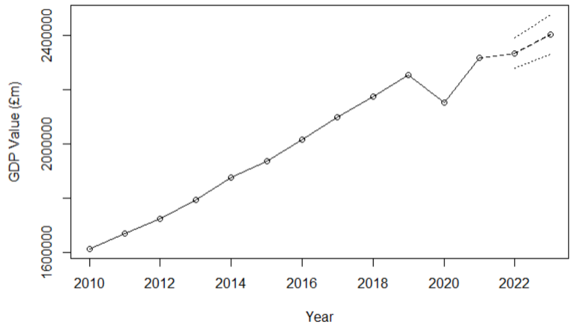
Figure 30 : Normality of the Residuals - QQ plots

Above plots illustrate that the quantiles of the residuals is close to the normal distribution in general.However, few deviations are noticable at the edge of the line. Hence, A statistical test has been performed to check the normality of the distribution. Shapiro-Wilk test of normality has a test statistic p-value is lesser than 0.01 and HO should be rejected. Hence, the residuals are not normally distributed. However, the reseduals do not indicate a pattern but distributed in randomly. Because, it is random part that cannot model. Hence, assume it has a normal distribution, but random.

Over fitting exercise was carried out by adding more parameters to evaluate whether the selected model is overfitting. As such, two parameter has added to our model as below. New model (pq31) has four parameters (ar1, ar2, ar3 and ma1) and the estimation for extra parameter is 0.1595, and -0.0225 which is very close to zero. Hence, adding a new parameter (ar3) is insignificant and made the model ineffective. Further, the AIC value for the new model (pq31) is 1692.17 which is slightly higher than the original model 'pq11'. Hence, It provide a sign that the selected model is a overfitted model.

Forecasting of the model

Forecasting for the model pp0201 (ARIMA(1,2,1)) as below.



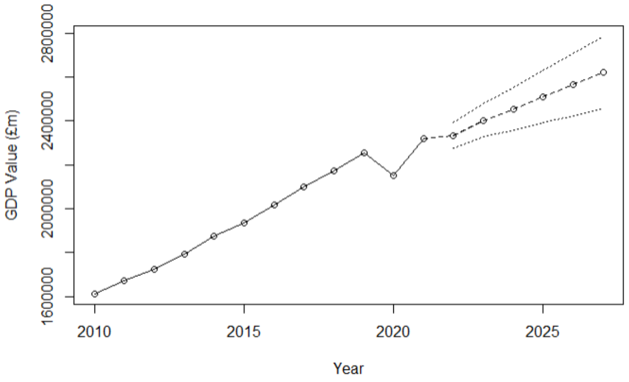
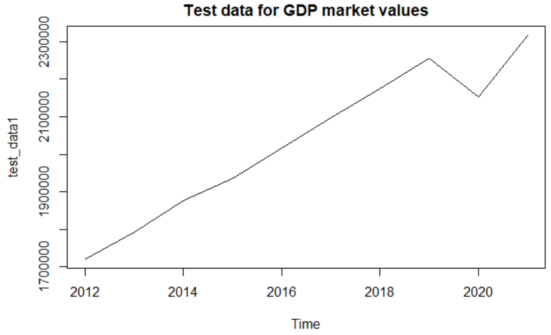


Figure 31 : GDP value for 1 years prediction

Figure 32 : GDP value for 4 years prediction

Forecasting was done for next 2 and 4 years respectively. It could examine that the forecasted values of the model (ARIMA (1,2,1) is similar to the pattern of the previous years irrespective the random drop in 2020. Hence, model was fitted to forecast the values. The time plot for next 4 years indicate that the interval of confidence increases as the period increases which leads predictions become less precise. Hence, forecasting for quite lengthy periods tend to have increased interval of confidence.

A forecast “error” is the difference between an observed value and its forecast. Accuracy of the model prediction depends on the length of the trained data. Hence, the data set used for this analysis has quite lengthy period of past data starting from 1948 to 2021. Train and test data was allocated based on the 80:20. In general, above two test results of forecast errors for the two models namely pq11 and pq21, provide lesser values for the pq11 model which is ARIMA(1,2,1). Hence, based on the analysis of forecast errors, it assume ARIMA(1,2,1) model out of the tested models.



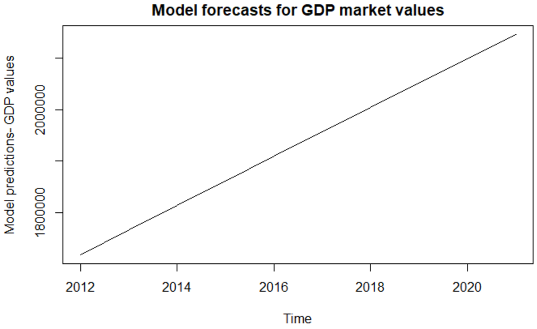
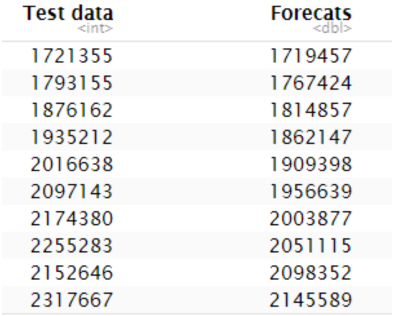


Figure 32 : Test Data

Figure 33: Model Predictions for test data

Above output illustrate the values for the test data and model predicted value for test data along with the forecasted values for each of year. As such, the predicted values of the fitted model is closer to the actual values. Further, below plots illustrate the patterns relevant to the above data.



Test data <int>	Forecats <dbl>
1721355	1719457
1793155	1767424
1876162	1814857
1935212	1862147
2016638	1909398
2097143	1956639
2174380	2003877
2255283	2051115
2152646	2098352
2317667	2145589

Figure 34 : Comparison of Predictions for the year 2020

Conclusion:

Time series for number of annual GDP at market prices ((£m) of UK was analysed in this study. As such, time series was convert to a stationary time series, develop list of possible models, model fitting and residual analysis was performed. Finally, ARIMA(2,2,1) was fitted to forecast the Gross Domestic Product at market prices. Later, time series was segregated to train and test data based on 80:20 proportion. Then, ARIMA(2,2,1) was tested for train and test data set and evaluate the prediction criteria. As such, the model follows the similar pattern to that of past data and predicted values closer to the test value set.

This model can be improved by application of some machine learning prediction algorithms to predict the values almost similar er to the test values. Further, this data set has monitory values and should be adjusted for inflation which has not captured in this model development. Hence, another model could be develop to track the inflationary impact over the GDP overtime. # predict.Arima: Forecast from ARIMA fits

Application of SARIMA to a monthly data set

Exploratory Data Analysis

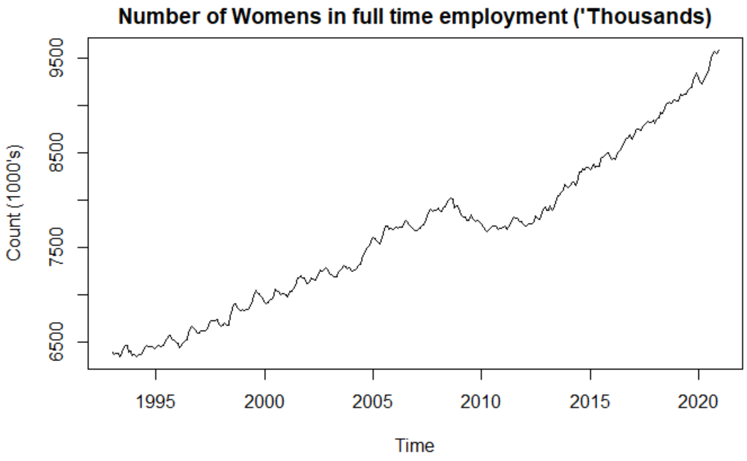
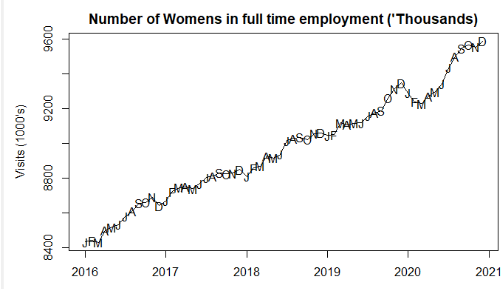


Figure 35 : Time series

Above Time plot in figure 35 illustrate followings; there is a increasing trend (upward trend) as the time goes up, the number of visits also growing up in general. The time series displays a some regular pattern. This provides a sense of seasonality. However, further analysis has been performed to study the seasonal effect in below.

The variance (size of employment count) is not increasing as the time (level of the series) increases. Hence, Box-Cox transformation has not been applied to this time series. All the peaks and drops are not in same size, as there is some viability. As such, it can assume that the Time series variance is not constant over the time.



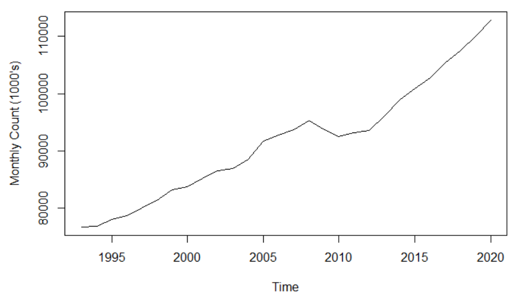


Figure 36: a portion of recent time slots

Figure 37: Aggregated data

Time series object has been plotted (Figure 36) for a portion of recent time slots to see a seasonal pattern of the time series. Above time plot indicates a regularly pattern of highs and lows within the year. Theres a slight drop in each January month and a increase in November and December months. As such, it provides a sense of seasonality component associated with this time series.

The aggregate (annual) data was plotted (Figure 37) to see the trend pattern of annual data and which indicates a upward trend as the time goes up the visits also growing up.

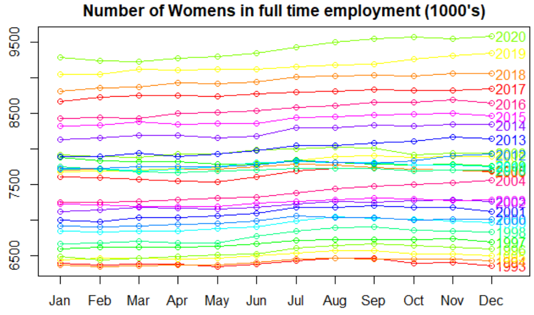


Figure 38: Seasonality plot slots

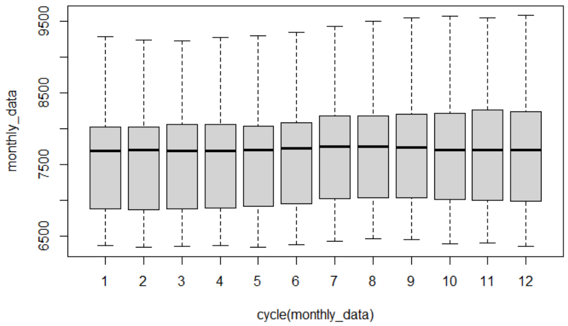


Figure 39: Boxplot

Above seasonal plot (Figure 38) indicate the underlying seasonal pattern more clearly. There is no massive ups and downs between months. However, gradual increase in December is visible at the latter part of the period of consideration. The Boxplot (Figure 39) is a form of plot enables the underlying seasonal pattern to be seen clearly. The above Boxplot indicates the positions in the cycle of four frequencies. There could observe a slight pattern when observing the median line of the Boxplots which at the beginning of the year tend to decrease. Then at the middle of the year which correspond to summer time there is a increase. Further, the range has been increase gradually from January to December.

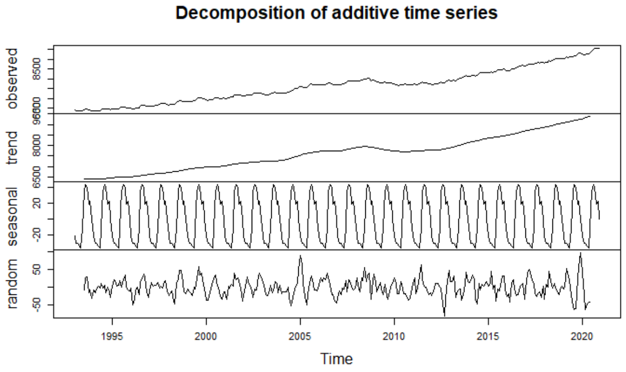


Figure 40: Additive model

Additive model was used as the time series variance (size of the visits) is not increasing as the time (level of the series) increases by (variance volatility is not increasing with the time). Decomposition plot illustrate the trend, seasonal and random component of the time series separately in a graphical manner as above. The underlying time series is a non-stationary time series as it has a trend or a seasonal component and requires differencing to transform it to a stationary time series.

Model fitting and Forecasting

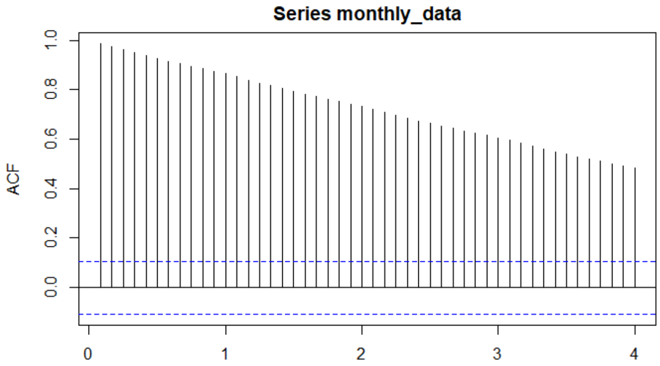
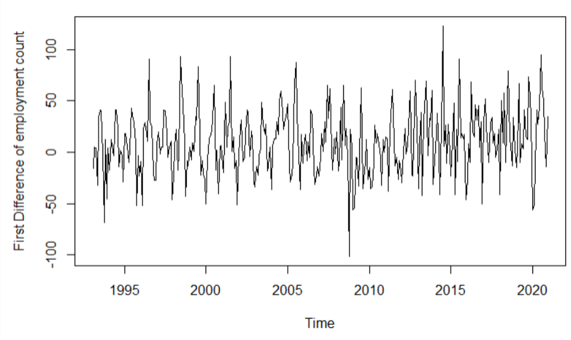


Figure 41: Correlogram of original TS

Above plot of ACF indicates significant positive many autocorrelations. The autocorrelations decreases slowly as the number of lags increases. The autocorrelations are gradually decreasing after the every peak as the lags increases. A combination of these effects could be seen, when data are both trended and seasonal. Hence, ACF indicates trended time series. In a stationary time series, the ACF will drop to zero relatively quickly as it doesn't depends on time, while the ACF of non-stationary data decreases slowly as in the above plot. Further, more than 5% of spikes are outside the bounds. As such, the above ACF indicates the attributes of non-stationary time series or not a white noise.

Further, it is also useful to quantify the evidence of non-stationarity via hypothesis testing. Hence, the Dickey -Fuller test and KPSS tests has been used. The p-value of above ADF test is 0.99, which is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that there the time series is not stationary. The p-value of the KPSS test is 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is not a stationary. It is apparent that the above time series is a non-stationary since it has a trend and seasonal component. Hence, it is required to convert to a stationary time series to construct a appropriate model for forecasting. Decompose the series into the components trend, seasonal effect, and residuals, and plot the decomposed series in below.



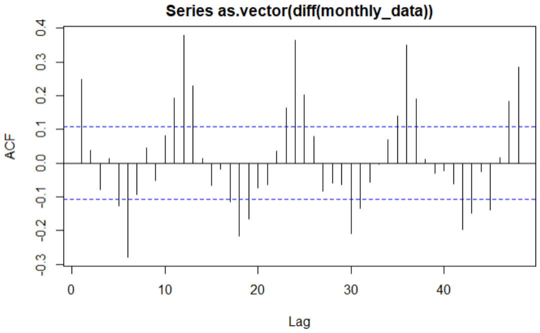


Figure 42: First differenced time plot

Figure 43: ACF after first difference

A time plot after the first difference is illustrate above. In this case, it is clear that the trend component has been eliminated from the series and values oscillate around a constant level. Then, seasonal differencing is required to covert the time series to a stationary one or a white noise. Above ACF plot indicates autocorrelation function for the first Difference. There are many significant spikes with a pattern. Hence, second difference is applied to the time series to remove the seasonal effect.

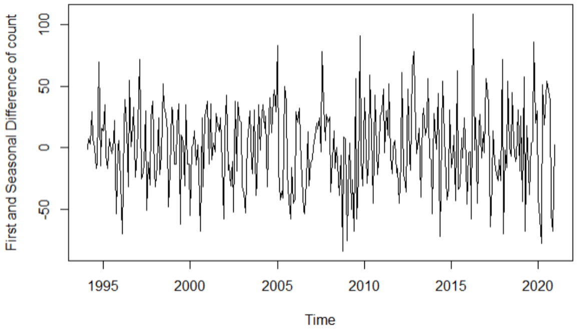


Figure 44 : First and Seasonal Difference

The above figure displays the time plot after application of both first difference and seasonal difference. As such, there is no trend component, since the values are oscillate around a constant level. Further, a random pattern is visible. Accordingly, both the trend and seasonal component has been eliminated from the time series. Further, the evidence of non-stationarity could be quantify via hypothesis testing. Thus, Dickey - Fuller test has been applied below.

The p-value of above ADF test is 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is a stationary. The p-value of above KPSS test is 0.1, which is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that there the time series is a stationary.

Model Specification

The following ACF, PACF and EACF plots have been analysed to determine the parameters initially.

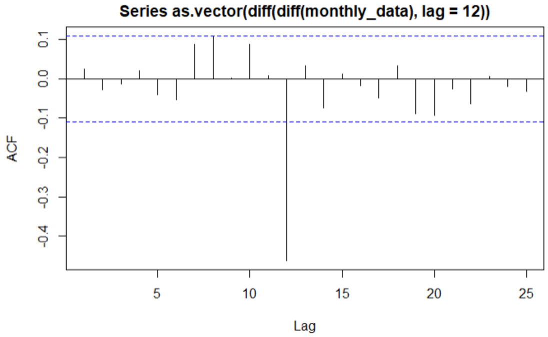


Figure 45: ACF

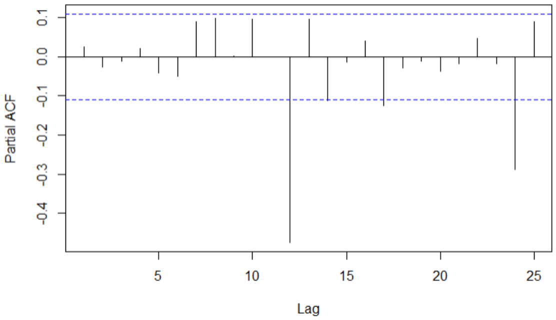
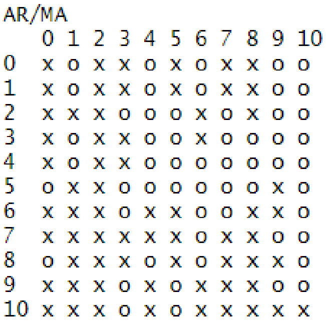


Figure 45: PACF

Above ACF function illustrate that approximately 95% of the spikes are within the interval of confidence. However, there is one significant autocorrelations are out from the interval of confidence. The spikes at lag 12 which is significantly different from zero and rest of all lags are zero or closer to zero. Hence, it could assume that the time series is a is a stationary. Above ACF plot has been used to determine moving average (MA) component of the model specification. As such, ACF illustrate both a non-seasonal moving average component (q) and seasonal moving average component (Q). Hence, it could assume 'q' as 0 or 1, and 'Q' as 0. The first difference (d) was set as 1.

PACF function indicates that very few significant autocorrelations are out from the interval of confidence. However, it can assume that the 95% is within the interval of confidence. The spikes at lag 12 and 24 is significantly different from zero and rest of all lags are zero or closer to zero. Hence, it could assume that the time series is a is a stationary. PACF plot illustrate autoregressive (AR) component of the of model specification. The time series in this analysis consist with both non-seasonal autoregressive component (p) and seasonal autoregressive component (Q). Significant Peaks tend to be four lags apart. Hence, by examining the above PACF plot,it could assume 'p' as 0 or 1, and 'P' as 1 or 2. The seasonal difference (D) was set as 1.

By examining the EACF plot (Figure 46) to determine a appropriate model is complicated as it does not have a clear visible pattern.



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

Figure 46: EACF

Parameter Estimation

As per the above output results,the lowest AIC value which is 3057.13 is provide by "m8" model and then the second lowest AIC value of 3057.23 for "m9". Hence, it can assume 'mpq1121' as the best model based on the AIC value and the second best as 'm8'. The attributes of the two models are describe below.

Further, BIC criteria has used to evaluate the models which has the lowest AIC values. As such, BIC criteria provides lowest value for "m8" which is 3078.016 which is slightly higher than the other model. Hence, both tests suggest the "m8" appears to be an more appropriate model for the time series that is seasonal ARIMA(0,1,1)×(1,1,2)12

Model attributes comparison; the lowest AIC value of 3057.13 is given by 'm8' model which has 4 parameters and the coefficients of parameters (ma1, sar1, sma1,sma2) is 0.0895, -0.7386, -0.1074 and -0.6654 respectively. Square error also quite small for all the parameters. The second lowest AIC value of 3057.23 is given by 'm9' which also has 4 parameters and the coefficients of parameters (ma1, ma2, sar1, sma1) is 0.0919, 0.014, 0.0284 and -0.8771 respectively.

As such, two models (m8 and m9) has been specified as below.

- **Best model (m8) out of the evaluated list of models : Seasonal ARIMA(0,1,1)×(1,1,2)12**
- **Second best model (m9) out of the evaluated list of models : Seasonal ARIMA(0,1,2)×(1,1,1)12**

Residual Analysis (Estimate of White noise)

Residuals should be a white noise which is a weaker condition of stationary. Accordingly analysis is performed for residuals as below for the above selected two models; Seasonal ARIMA(0,1,1)×(1,1,2)12 and Seasonal ARIMA(0,1,2)×(1,1,1)12

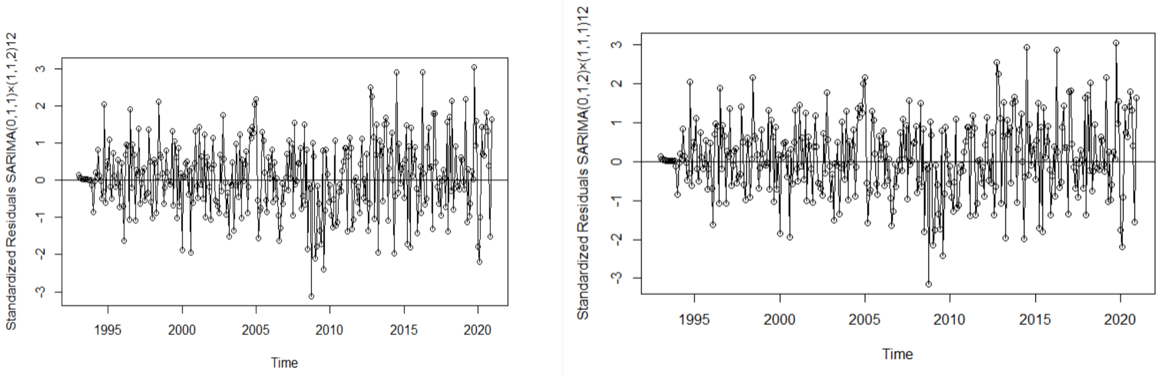


Figure 47 : Resedual Analysis

The above residual plots illustrate that the majority of residuals are oscillate around zero and approximately 95% of data fall between +2 & -2. As residuals should be a stationary which is independent and identically distributed. Hence, ACF of residuals should be close to zero and 95% of lags should be inside the interval of confidence.

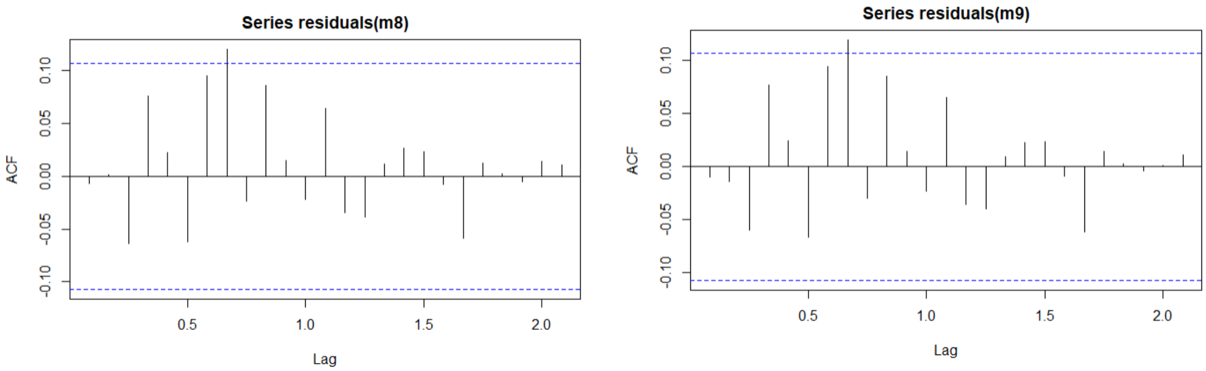


Figure 48 : ACF plots for Reseduals

ACF function for residuals indicates that only one or two autocorrelation are out from the interval of confidence. Hence, that the 95% is within the interval of confidence and are approximately zero or closer to zero. As such, it appears to be no significant autocorrelation in the residuals and the residuals are uncorrelated. The p-value of above Box-Pierce test and the Ljung-Box test are not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest for Residuals or error terms are uncorrelated.

The p-value of both ADF test are below 0.01, which is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H1) and conclude that the time series is a stationary. This confirms the residuals are stationary or white noise. Above plots illustrate that the quantiles of the residuals is close to the normal distribution in general.However, few deviations are noticable at the edge of the line. Hence, A statistical test has been performed to chack the normality of the distribution. Shapiro-Wilk test of normality has a test statistic p-value is 0.8217 and HO should not be rejected. Hence, the residuals are normally distributed.

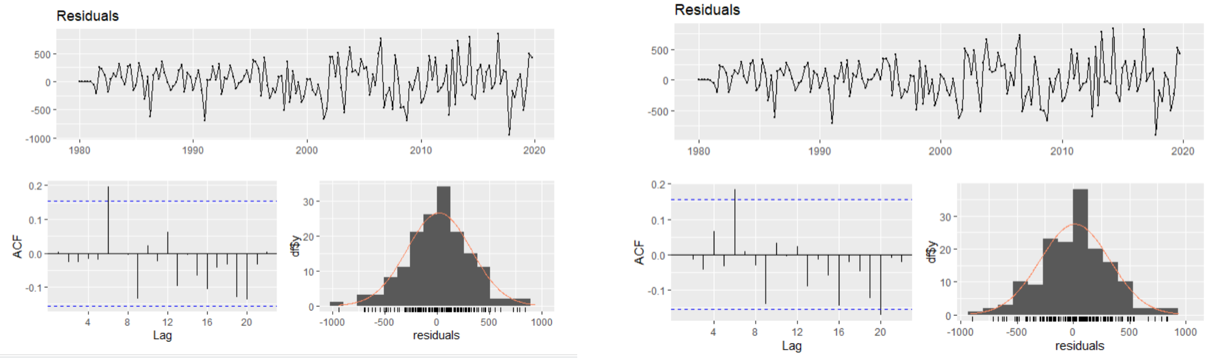
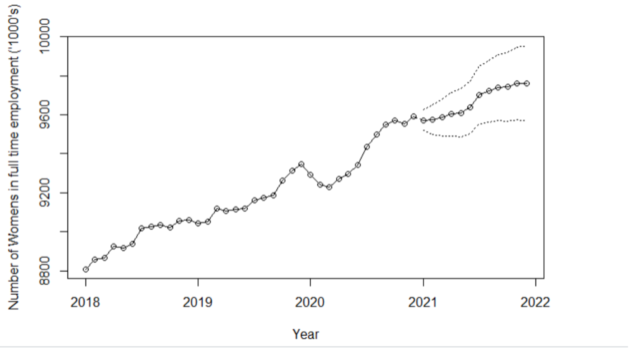


Figure 49 : Summary of resedual analysis

Residuals analysis (Figure 49) illustrate that the residuals are stationary/white noise (weaker conditions), uncorrelated (correlation=0, independent) and normally distributed for both the models.

Forecasting of the model

Forecasting for the model pp0201 which is (Seasonal ARIMA(0,1,1)×(1,1,2)12) is as below.



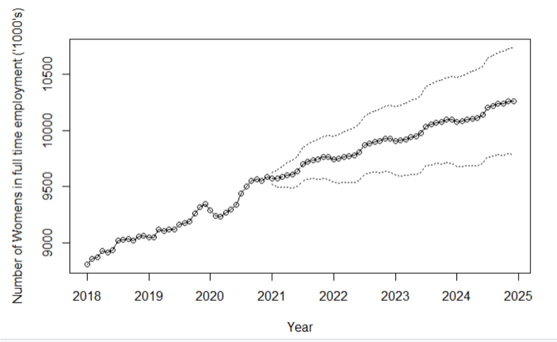


Figure 50 : Forecast for 1 year

Figure 51 : Forecast for 4 years

Seasonal ARIMA(0,1,1)×(1,1,2)12 model was used to forecast for next 12 and 48 months respectively. It could examine that the forecasted values of the model (Seasonal ARIMA(0,1,1)×(1,1,2)12 is similar to the pattern of the previous months of the year. Hence, model is appropriate and fitted to forecast the values of the time series. The time plot for next 48 months indicate that the interval of confidence increases as the period increases which leads predictions become less precise. Hence, forecasting for quite lengthy periods tend to have increased interval of confidence.

A forecast "error" is the difference between an observed value and its forecast. Accuracy of the model prediction depends on the length of the trained data. Hence, the data set used for this analysis has quite lengthy period of past data starting from Qtr1 of 1980 to Qtr 4 of 2019. Test results is related to the forecast errors of the two models namely m8 and the m9. Lesser values for forecast error is given by the m9 model which is Seasonal ARIMA(0,1,2)×(1,1,1)12. Hence, based on the analysis of forecast errors, it assume Seasonal ARIMA(0,1,2)×(1,1,1)12 as the best model out of the tested models.

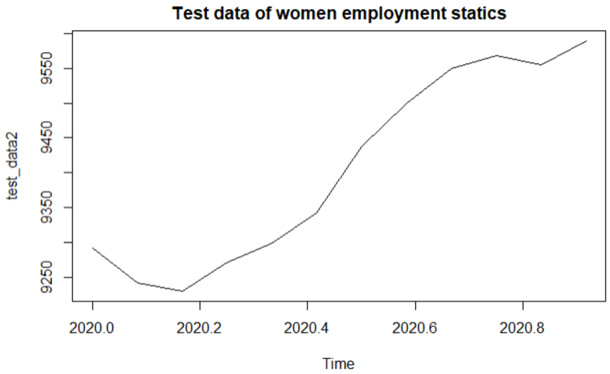


Figure 52: Test data

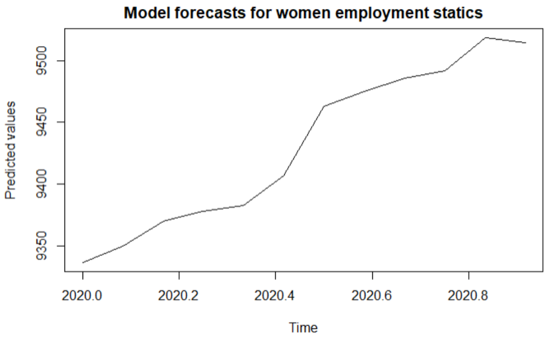
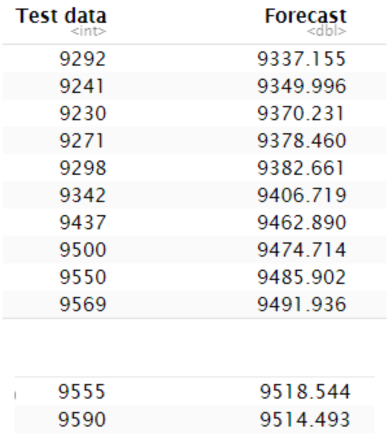


Figure 53: Predictions for test data

Below Figure 54, output illustrate the values of test data and predicted value of Seasonal ARIMA(0,1,2)×(1,1,1)12 model for the test data on the months of year 2020. As such, it was observed that the predicted values of the model follows similar pattern to the previous data (Figure 52 and 53). Further, values predicted of the fitted model is closer to the actual values. Further, below plots illustrate the patterns relevant to the above data.



Test data <int>	Forecast <dbl>
9292	9337.155
9241	9349.996
9230	9370.231
9271	9378.460
9298	9382.661
9342	9406.719
9437	9462.890
9500	9474.714
9550	9485.902
9569	9491.936
9555	9518.544
9590	9514.493

Figure 54: Comparrison of test data and predictions

Conclusion:

Time series for Womens in full time employment statistics in UK was analysed in this study. As such, time series was convert to a stationary time series, develop list of possible models, model fitting and , residual analysis was performed. Finally, Seasonal ARIMA(0,1,2)×(1,1,1)12 was fitted to forecast the number of overseas visits to UK. Then, apply that model to train and test data set of the same time series and evaluate the prediction criteria. As such, the model follows the similar pattern to that of past data and predict values closer to the test value set.

This model can be improved by application of some machine learning prediction algorithms to predict the values almost similar to the test values. Further, this data set is related pre-covid situation. However, there may be variations during covid seasons and afterwards. The data set used in this analysis is pre-Covid period. Hence, another model could be develop to track the during_Covid and post_Covid forecasts.