**UNIVERSITY OF SOUTH WALES**

**DATA MINING AND STATISTICAL FORECASTING**

**COURSE CODE: MS4S09**

**COURSE WORK 2**

**TIME SERIES ANALYSIS**

**SUBMITTED BY**

**NAME: NISHA MUNDANTHARA SURESH BABU**

**STUDENT ID: 30042535**

**ARMA**

**Problem Statement:**

Heavy fuel oils are primarily used as a marine fuel. So, the price change of that has great importance in the economy of that country. The price of this fuel can fluctuate greatly but the change of weekly prices seems to be stationary.

The aim of the course work is to perform time-series analysis to investigate how well modeling could forecast the weekly price change of heavy fuel oil in Italy.

**Data**

The data on weekly price change for heavy fuel oil have been collected from the website [*https://data.world/rafabelokurows/weekly-fuel-prices-in-italy*](https://data.world/rafabelokurows/weekly-fuel-prices-in-italy) for the period from January 2005 to August 2021.

**Exploratory Data Analysis**

Summary statistics have been performed and the result is in Appendix A. There are 830 observations with 8 columns. Since we are modeling the price change, we are only considering the columns ‘survey\_date’ and ‘CHANGE’. The mean change is 0.3437. There are no null values. The plots in Fig1.1 (b) shows neither trend nor seasonality, but some significant dips(outliers) can be seen in some weeks of 2009,2013 and 2019. The decomposed plots (see Appendix A) also show a similar view. This should be further examined using box plots. In the box plot from Fig 1.1(a) we can see that there is a significant number of outliers appearing in the CHANGE column which may seriously affect the forecasting. Therefore, the tsoutliers function is used to detect and remove the outlier in the time series.

Graphical user interface

Description automatically generated

Figure 1.1(a) Outliers in Price change (b) Weekly fuel price change from 2005 to 2021

**Stationarity of the time series**

In order to test the stationarity of the time series, the Augmented Dickey-Fuller Test is performed. The Hypothesis for that is:

|  |
| --- |
| H0 : The time series is non-stationary  H1 : The time series is stationary |

Here p-value is 0.01 which is < .05, so we reject the null hypothesis and conclude that the time series is stationary. The correlogram in Fig1.2 also supports the stationarity because of the fast decaying of the series.

Chart

Description automatically generated

Figure 1.2 ACF for fuel price change

**Model Used**

Autoregressive (AR) models, coupled with moving average (MA) models to form a general and useful class of time series models known as autoregressive moving average ARMA (p, q) models. ARMA models are very useful when the time series is stationary. Since our data is stationary we used an ARMA model.

**Model Selection**

The fig 1.4 indicates that the auto-correlations have three spikes which indicate a MA (3) behavior. At the same time, fig1.3(a) reveals that the partial autocorrelations have also three spikes by which it is inferred that an auto-regressive of order three (AR3) is appropriate for the time-series data. We can also combine AR and MA to get an ARMA(3,3) model. Also the EACF matrix in fig 1.3(b) shows an ARMA(1,1). All the models are calculated and their efficiencies are compared using Information Criterions and the results are shown in Table 1.1. The least AIC and BIC are for p=1 and q=1. So we select the model ARMA(1,1) as the best.

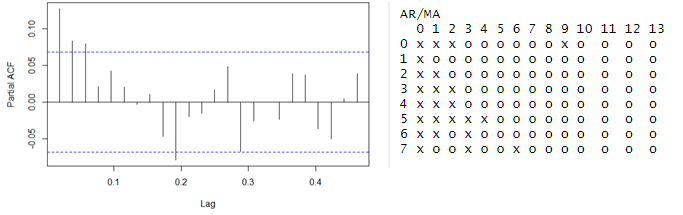


Figure 1.3(a) PACF for fuel price change (b) EACF matrix

|  |  |  |
| --- | --- | --- |
| Model | AIC | BIC |
| MA(3) | 6591.8 | 6617.408 |
| AR(3) | 6589.54 | 6615.147 |
| ARMA(3,3) | 6592.2 | 6631.974 |
| ARMA(1,1) | 6587.13 | 6608.02 |

*Table 1.1 Comparisons of AIC and BIC*

The time series has been checked for overfitting by adding another MA and AR parameters and checked for accuracy yields higher AIC values (See Appendix A). Hence concluded with the model ARIMA(1,0,1). The selected model can be represented as

Xt = φ1Xt−1 + et − θ1et−1

The estimated parameters are: φ = 0.7499, θ = -0.6346, so the final model is:

Xt = 0.7499 Xt−1 + et + 0.6346et−1

**Residual Analysis**

For a good model, the residual should follow the features of white noise. So, a residual analysis is necessary to find this. The plot for residual analysis is given in Fig 1.4. From the normal Q-Q plot, we can see that we almost have a straight line, which suggests no systematic departure from normality.The residual plot looks stationary because there's no obvious structure in the residuals. The correlogram on the bottom left has a few spikes outside the boundary but we can consider it as the 5% data which deviates from the usual structure. So, we can conclude that there is no autocorrelation in the residuals, and so they are effectively white noise. We can further guarantee this using the box test. The hypothesis for this is as follows:

|  |
| --- |
| H0 : the data are sample from an iid sequence  H1 : the data are not the sample from an iid sequence. |

This leads to a p-value of 0.8032 and 0.3881 or the Box-Pierce test and the Ljung-Box test, respectively, so we have no evidence to reject the null hypothesis that the error terms are uncorrelated. We also checked the stationarity of the residuals using Dickey-Fuller test, this results in a p-value 0.01 also shows stationarity.

Graphical user interface, diagram, engineering drawing

Description automatically generated

Figure 1.4 Residual plots

**Forecasting**

#### The data is divided into two sections: training and testing. The data from year 1998 to 2017 as training set and 2018 to 2021 as test set. We fit train data into our model and compare the results to actual data to determine how accurate our model is. If our model's accuracy is high enough, we can forecast the future. The correctness of the fitted model is measured by MAE, and RMSE values. The Mean Absolute Error (MAE) is defined as the average of the absolute difference between forecasted and true values. Root Mean Squared Error (RMSE) is a standard deviation of prediction errors. It indicates how spread out the data is around the line of best fit. The model with least measure generally indicates the perfect model for prediction.

Table 1.2 shows the RMSE values for the predicted values for all the three methods. ARIMA(3,0,3) has the least value of RMSE so we are concluding that the model ARIMA(3,0,3) is the best model for forecasting. The forecasted values and test values are shown in Appendix A.

|  |  |  |
| --- | --- | --- |
| Model | RMSE | MAE |
| ARMA(1,0,1) | 3.446028 | 3.446028 |
| ARMA(3,0,0) | 3.443194 | 3.443194 |
| ARMA(0,0,3) | 3.440426 | 3.440426 |
| ARMA(3,0,3) | 3.429901 | 3.429901 |

Table 1.2 RMSE values of models

Now we have trained and tested our model we can predict the future values as well. Figure 1.5 shows the forecasts and 95% forecast limits for a lead time of two years for the ARMA(3,3) model that we fit. Though the forecasted values are inside the 95% confidence interval, the selected model doesn’t seem to perform well.

Chart

Description automatically generated

Figure 1.5 Price change forecasting

**Conclusion**

This paper examines univariate time series ARMA forecasting method for fuel price change of heavy oil in Italy. The time series has tested for stationarity and the selection of the best model has been conducted based on information criterion. ARIMA(1,0,1) model has been selected as the best fit model. The data is further divided into training and test set based on the general ratio 80:20 and trained for the model ARIMA(1,0,1). But the least RMSE value has been found to be for the ARIMA(3,0,3) and predicted future values. The predicted values show severe deviation from the original data. This makes the model less suitable for forecasting. In future work, we can add additional variables to get a good forecasting model.

**ARIMA**

**Problem Statement:**

The National Statistics Retail Sales ex-fuel is a measure of changes in sales in the British retail sector excluding fuel. The primary goal of the Retail Sales Index (RSI) is to produce a short-term measure of changes in the volume and value of goods sold by retail businesses in the United Kingdom, providing a timely indicator of economic performance and consumer spending strength.

The course work aims to build an ARMA model and forecast the future values of RSI values of the UK retail sector excluding fuel.

**Data**

The dataset is downloaded from the UK government’s office of national statistics website [*https://www.ons.gov.uk/businessindustryandtrade/retailindustry/timeseries/j467/drsi*](https://www.ons.gov.uk/businessindustryandtrade/retailindustry/timeseries/j467/drsi). It contains quarterly RSI values ranging from the year 1988 to 2019.

**Exploratory Data Analysis**

Summary statistics have been performed and the result is shown in Appendix B. There are 128 observations with 2 columns ( Year and RSI). The Year column range from 1988 Quarter1 to 2019 Quarter4. The mean value of RSI is 71.33. There do not appear to be a significant number of outliers and there are no missing values. Therefore, no further data cleaning is required.

The plots in Fig2.1(a) show a clear positive trend. In the decomposed plots from Fig2.1(b) we can again see the trend and we can also observe the estimation of the random component depicted under the remainder

Chart

Description automatically generated

Figure 2.1(a) Line plot of RSI time series (b) Decomposition of time series

**Stationarity of the time series**

To test the stationarity of the time series the Augmented Dickey-Fuller Test was performed. The Hypothesis for that is:

|  |
| --- |
| H0 : The time series is non-stationary  H1 : The time series is stationary |

Here p-value is 0.8177 which is > .05, so we accept the null hypothesis and conclude time series Is not stationary. Further evidence is provided by the correlogram in Fig 2.2. Since there is no pattern in the correlogram, we can say there is no seasonality. But it clearly shows a trend.

Chart, bar chart

Description automatically generated

Figure 2.2 Correlogram for the time series

**Model Selection**

Since the series has only trend and not seasonality an ARIMA model will be suitable. First, we will use log transformation of the series in order to get a constant variance and check the stationarity again. But still, the data is non-stationary. So, we will do a first-order differencing on the series to remove the linear trend.

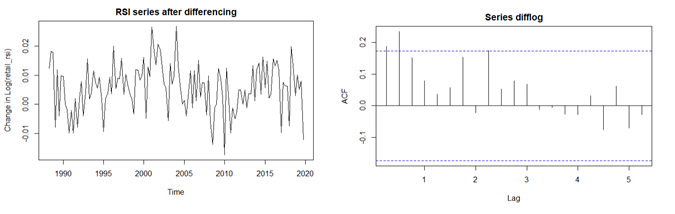


Figure 2.3 (a): RSI series after differencing (b) ACF after first order differencing

Fig2.3(a) shows the plot after differencing of log (RSI). It looks like a stationary one and the ACF after differencing decays fast to zero. We can further confirm stationarity by the Augmented Dickey-Fuller Test results. This results in a p-value 0.04729 which is < .05, so we reject the null hypothesis and conclude the time series Is stationary.

**Model Fitting**

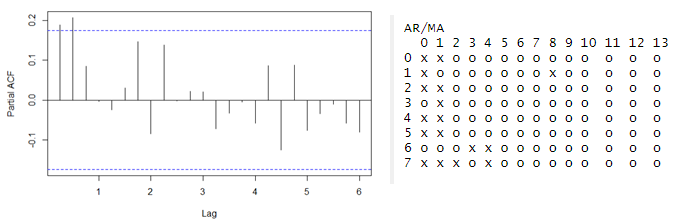
An ARIMA(p,d,q) model is a combination of an AR model, MA model, and differencing where, p and q are the order of the AR and MA part respectively, and d is the degree of non-seasonal differencing. To find out the p,q, and d values we usually analyse ACF and PACF plots. The bars that fall above or below the dashed line represent statistically significant (*p* < 0.05) autocorrelation. The ACF (Fig 2.3(b) ) plot suggest a MA(2), the PACF (Fig 2.5(a) ) suggests an AR(2) , and the EACF matrix (Fig 2.5(b)) suggests an ARMA(1,1). The model incorporates a first-order difference (*d* = 1) to eliminate trend and induce stationarity. 

Figure 2.4 (a) PACF after differencing (b) EACF Matrix

To compare the models, I make a comparison of the information criteria (Table 2.1). From the following table, we can see that the ARIMA(1,1,1) model is much better than that of the other 2 models. Thus, we will consider this as our potential final model.

|  |  |  |
| --- | --- | --- |
| Model | AIC | BIC |
| ARIMA(0,1,2) | -834.51 | -823.9786 |
| ARIMA(2,1,0) | -850.33 | -839.7995 |
| ARIMA(1,1,1) | -857.72 | -847.192 |

Table 2.1 Comparisons of Information Criteria

The least AIC and BIC are for p=1 and q=1. So we select the model ARIMA(1,1,1).

The time series has been checked for overfitting by adding another MA and AR parameters and checked for accuracy yields higher AIC values (See Appendix B). Hence concluded with the model ARIMA(1,1,1). The selected model can be represented as

Xt = φ1Xt−1 + et − θ1et−1

The estimated parameters are: φ = 0.9682, θ = -0.7895, so the final model is:

Xt = 0.7499 Xt−1 ++ et + -0.7895et−1

**Residual Analysis**

The residual plots are shown in Fig. [5](https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01235-8#Fig4). There is no obvious pattern or significant autocorrelation in the residuals, and they are normally distributed. The *p*-value for the Ljung-Box test for white noise is 0.90 at 24 lags. As the null hypothesis for the Ljung-Box test is that there is no significant autocorrelation, we do not reject the null and our chosen model has a good fit.

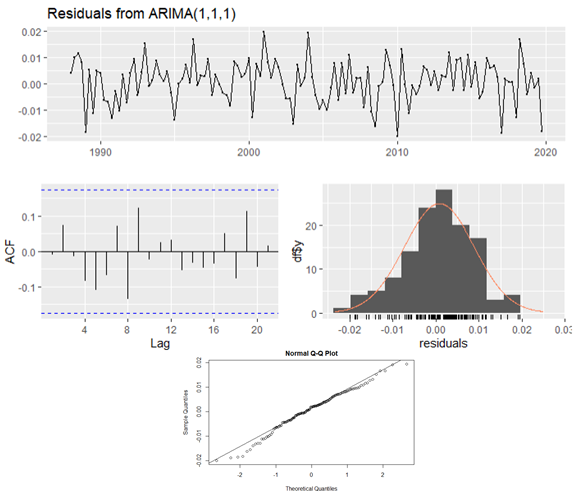


Figure 2.5 Residual Plot

**Forecasting**

The data is split into two sets. The data from year 1998 to 2017 as training and 2017 to 2019 as test set where the training data is used to estimate any parameters of a forecasting method and the test data is used to evaluate its accuracy. Figure 2.8 shows prediction using our training dataset. The black line is the actual data, blue line is the predicted data and the orange line is the expected error range. Our model seems to predict the values correctly except in the end of 2020, this may be the effect of COVID-19.

Table 2.2 shows the RMSE values for the predicted values for all three methods. Since ARIMA(1,1,1) has the least value of RMSE, it implies that this model is the best model for forecasting. So ARIMA(1,1,1) is a good fit for forecasting as well. The forecasted values and test values are shown in Appendix B.

|  |  |  |
| --- | --- | --- |
| Model | RMSE | MAE |
| ARIMA(1,1,1) | 03607534 | 0.03305741 |
| ARIMA(1,1,0) | 0.04985682 | 0.04542249 |
| ARIMA(0,1,1) | 0.05144211 | 0.04709806 |
| ARIMA(0,1,2) | 0.04857942 | 0.04386746 |
| ARIMA(2,1,0) | 0.03983798 | 0.03569467 |

*Table 2.2 RMSE values of models*

Chart, line chart

Description automatically generated

Figure 2.7: Forecasting with ARIMA(1,1,1)

Figure 2.7 shows the forecasts and 95% forecast limits for a lead time of two years for the ARIMA(1, 1, 1) model that we fit. The last three years of observed data are also shown. The final model has managed to capture the trendline and fit inside the confidence interval.

**Conclusion**

Overall this session investigates the univariate time series ARIMA forecasting method for Retail Sales Index (RSI) of all the businesses except fuel in the UK. The time series exhibits a linear trend and also the ACF reduces to zero very slowly indicating a non-stationary model. A log transformation was conducted to stabilize the variance and perform a first-order differencing to remove the trend. Various values for p,q were identified based on the ACF, PACF and EACF values. These ARIMA models were tested based on the information criterion and selected the best model. ARIMA(1,1,1) model has been selected as the best fit model. The data is further divided into training and test set based on the general ratio 80:20 and trained for the model ARIMA(1,1,1). This model shows the least RMSE value also. Eight quarters of future data were predicted using the model selected. The predicted model was able to catch the trend line. So, growth in the RSI values can be expected in the future years as well. This is a good sign for the UK economy.

**SARIMA**

**Problem Statement:**

The rise in international travel has had a significant impact on employment, business and growth in tourism and related industries in many countries. The course work aims to perform a model selection strategy for overseas travel to the UK that accounts for three decades of overseas travel data.

**Data**

The dataset is downloaded from the UK government’s office of national statistics website [*https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/timeseries/gmaa/ott*](https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/timeseries/gmaa/ott). It contains monthly values of overseas visitors to the UK ranging from the year 1980 to 2020. All visits are in thousands.

**Exploratory Data Analysis**

Summary statistics have been performed and the result is shown Appendix C.There are 480 observations with 2090 thousands average visitors per month. There do not appear to be a significant number of outliers and there are no missing values. Therefore, no further data cleaning is required.

The plots in Fig3.1show a clear positive trend. From fig3.2(a) We can see that there is a seasonal variation in the time series within a year. Boxplot reveals that OS visits is higher in months July and August and lower in the month of February, indicating seasonality with an apparent cycle of 12 months. Fig 3.2(b) explains this more clearly.

Chart

Description automatically generated

Figure 3.1Plot for OS visitors to the UK

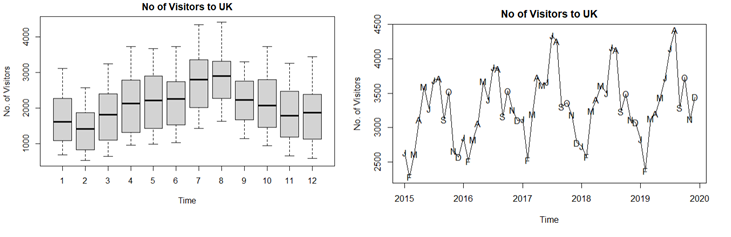


Figure 3.2(a) Box plot (b)Month wise visitors to the UK

**Stationarity of the time series**

Since the time series shows clear trend and seasonality, we can conclude that the time series is not stationary. The correlogram in Fig 3.3 gives further evidence to the non-stationarity because of the slow decaying of the series with seasonal effect.

Chart, histogram

Description automatically generated

Figure 3.3 Correlogram for the no. of visitors to the UK

**Model Selection**

Since the series has trend with seasonality a SARIMA model will be suitable. We will do a differencing on the series to remove the trend.

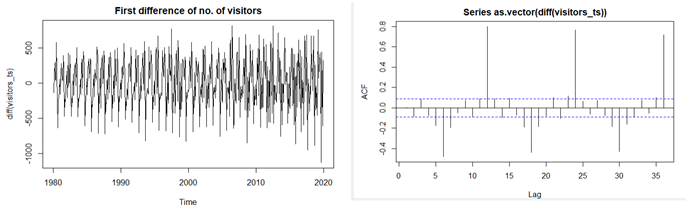


Figure 3.4 (a) Plot for first difference (b) ACF after first difference

The plot (Fig 3.4(a) looks like a no trend in this. The correlogram in Fig 3.4(b) decays faster. We observe from the ACF that the data has seasonality after every 12 months. Hence taking seasonal differences on the data would be a good idea. So now we will do differencing to remove the seasonality.

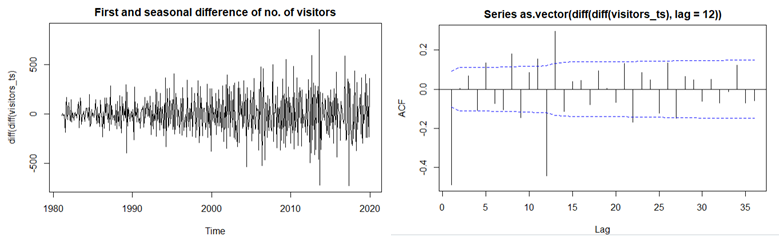


Figure 3.5(a) Plot after trend and seasonality removal (b) ACF

After the seasonal differencing note that the data has finally become stationary and ARIMA modeling can be done. The correlogram decays faster without any seasonal pattern. Also, the Augmented Dickey-Fuller Test results in a p-value 0.01 which is < .05, so we reject the null hypothesis and conclude the time series Is stationary.

**Model Fitting**

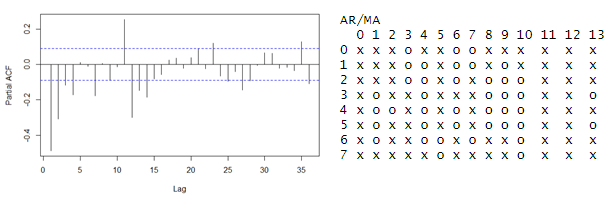


Figure 3.6(a) PACF (b) EACF matrix

From the ACF and PACF plot, we can say that it’s an ARIMA(p,d,q) X (P,D,Q) model. From Fig 3.5(b) we can see that there is 1 spike at the beginning of the ACF and there is another spike at 12th lag after that all the lags converge to zero. So we can take q=1 and Q=1. Likewise, From Fig 3.6(a) we can see that are 4 spikes at the beginning of the ACF and there is another spike at 13th lag after that all the lags converge to zero. So we can take p=3 (since the third lag is almost zero ) or 4 and Q=1. But considering the EACF matrix we can take the combination p=3 and q=1 as well. We try different values of p,q and P and Q and compare the BIC of each model.

The following table gives the comparisons of these 5 models.

|  |  |  |
| --- | --- | --- |
| Model | AIC | BIC |
| ARIMA(0,1,1)(0,1,1) | 5927.47 | 5941.91 |
| ARIMA(3,1,0)(1,1,0) | 6006.8 | 6029.529 |
| ARIMA(4,1,0)(1,1,0) | 6001.37 | 6028.249 |
| ARIMA(3,1,1)(1,1,1) | 5930.41 | 5961.43 |
| ARIMA(4,1,1)(1,1,1) | 5931.23 | 5966.401 |

Table 3.1 AIC and BIC values of models

The least AIC and BIC are for p=0,P=0, q=1, Q=1 d=1 and D=1 and seasonal component 12. So we select the model ARIMA(0,1,1)(0,1,1)12. The time series has been checked for overfitting by adding another MA and AR parameters and checked for accuracy yields higher AIC values (See Appendix C). Hence concluded with the model ARIMA(0,1,1)(0,1,1)12.

**Residual Analysis**

A residual analysis has been performed to find whether the residuals are the features of white noise or not. The plot for residual analysis is given in Fig 1.6.

The residual plots are shown in Fig. 3.7. There is no obvious pattern or significant autocorrelation in the residuals, and they are normally distributed. The *p*-value for the Box- Pierce test for white noise is 0.9108. As the null hypothesis for the Box test is that there is no significant autocorrelation, we do not reject the null and our chosen model has a good fit. Along with the Ljung Test, Dickey-Fuller Test was used to test whether the residuals were stationary. As the test showed that the residuals were stationary at p=.01 (Alternate hypothesis: Stationary).

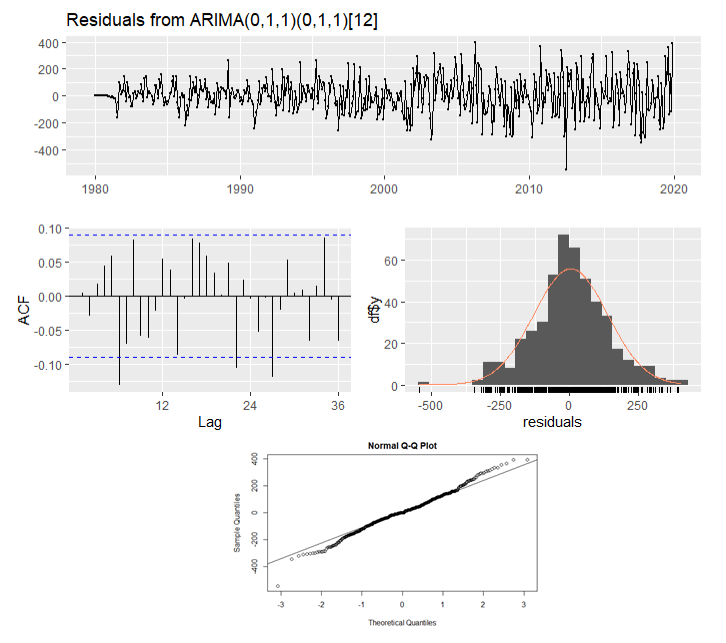


Figure 3.7 Residual Analysis

Forecasting

The data is split into two sets. The data from year 1980 to 2018 as training and 2019 to 2020 as test set where the training data is used to estimate any parameters of a forecasting method and the test data is used to evaluate its accuracy.

|  |  |  |
| --- | --- | --- |
| Model | RMSE | MAE |
| ARIMA(0,1,1)(0,1,1) | 198.8301 | 158.6075 |
| ARIMA(3,1,0)(1,1,0) | 220.8475 | 171.8803 |
| ARIMA(4,1,0)(1,1,0) | 210.454 | 154.8656 |
| ARIMA(3,1,1)(1,1,1) | 223.4257 | 163.7041 |
| ARIMA(4,1,1)(1,1,1) | 189.3475 | 149.4522 |

Table 3.2 RMSE values of models

Table 3.2 shows the MAE and RMSE values for the predicted values for all the three methods. Since ARIMA(4,1,1)(1,1,1) has the least value of MAE and RMSE, it implies that this model is the best model forecasting. The forecasted values and test values are shown in Appendix C. Thus after fitting the ARMA(4,1,1)(1,1,1) on the data I predict the no. of visitors to the UK for the next 24 months. Figure 3.8 shows the forecasts and 95% forecast limits for a lead time of two years.. The last three years of observed data are also shown. Our model managed to capture the trendline and seasonality.

Chart

Description automatically generated with medium confidence

Figure 3.8 Forecasting using ARIMA(4,1,1)(1,1,1)

Conclusion

In general, this session explores univariate time series SARIMA forecasting method for number of overseas visitors to the UK. The time series exhibits a linear trend and monthly seasonality suggested a non-stationary model. A first order differencing was performed to remove the trend as well as a seasonal differencing was done to remove the seasonality. Various values for p,q were identified based on the ACF,PACF and EACF values. These SARIMA models were tested based on the information criterion and selected the best model. ARIMA(0,1,1) (0,1,1) model was selected as the best fit model. The data is further divided into training and test set based on the general ratio 80:20 and trained for different SARIMA models The model ARIMA(4,1,1) (1,1,1) showed RMSE value less than the final model . 24 months of future data were predicted using the model selected. The predicted model was able to catch the trend line and seasonality. So, the increase in the number of overseas visitors with a seasonal effect is expected in the future years as well.

**Appendix A: ARMA**

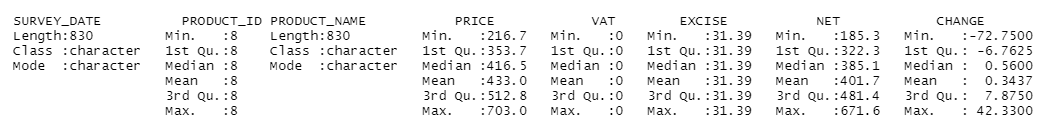
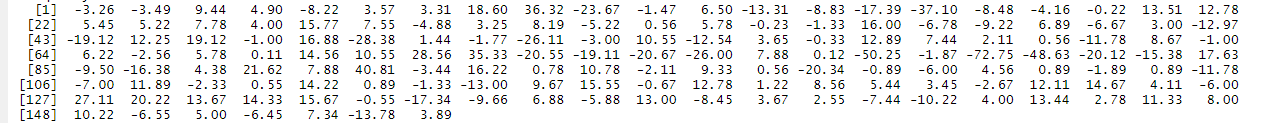


Figure 1.1 Summary Statistics

Diagram

Description automatically generated

Figure 1.2 Decomposed plot for fuel price change



Calendar

Description automatically generated with medium confidence

Figure 1.3 Test values vs Predicted Values

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1.4 Result of overfitting

**Appendix B: ARIMA**

Text

Description automatically generated

Figure 2.1: Summary Statistics

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Figure 2.2 Test values vs Predicted Values

Text

Description automatically generated with medium confidence

Figure 2.3 Result of overfitting

**Appendix C: SARIMA**

Text

Description automatically generated

Figure 3.1 Summary Statistics

Text

Description automatically generated

A picture containing table

Description automatically generated

Figure 3.2 Test values vs Predicted Values

Text

Description automatically generated with medium confidence

Figure 3.3 Result of overfitting