**Data Exploration of Real-World**

**Time Series**

Series 22 (D449), Series 50 (D478) & Series 78 (D506)

1. **Executive Summary**

The report discusses the data exploration process of three time series; Series 22, Series 50, and Series 78, at a descriptive stage where the data was analysed based on daily frequency up until quarterly frequency. Since the frequency of the original time series (daily) was extremely high it was difficult to visually identify any patterns in the data solely based on this. Henceforth, the data was aggregated at various levels such as weekly, monthly, and quarterly. Series 22 and 78 used the mean or average values while Series 50 used the sum or total values for these aggregates.

After exploring the graphical interpretation of all the time series, we have chosen an aggregate level to be the focus for the next stages of the analysis. This process was done based on the presence of seasonality identified through the seasonal plots from the seasplot function in R. On top of that, several types of decompositions were performed on the time series to identify any underlying or hidden seasonal patterns. The decompositions also helped in building up and strengthening our preferences for choosing an aggregate level. For Series 22 the aggregate level chosen is weekly while for Series 50 and 78, it was the monthly aggregate. The first two aggregate time series contain seasonality while the last aggregate time series lack this property. Both seasonal and non-seasonal time series were chosen to be analysed to see if there would be different outcomes in terms of how the analysis would progress.

Next, we performed two established stationarity tests, KPSS and ADF, on each of the time series to determine their state of stationarity. None of these was found to be stationary at the initial level. A differencing process of these time series would aid in transforming the data into a stationary state. Series 22 and 78 only required a first level of differencing to achieve a stationary state for both tests while Series 50 was a little bit more complicated as at the first level of differencing, only the KPSS test proved its stationarity, but following the suggestion of the nsdiffs and ndiffs functions available in R, we have decided to go on with this first level of differencing data for the next stage of analysis.

As our time series are all stationary, we were then able to conduct the autocorrelation function (ACF) and partial autocorrelation function (PACF) analyses. This is done via the tsdisplay function in R that will provide plots for both ACF and PACF as well as the differenced time series plot. From these plots, we were then able to create a few possible fits of ARIMA and even SARIMA models for forecasting. In the case of Series 22, we fortunately found the best fit for the model on the first try which was ARIMA(1,1,0). However, for the other series, we could not produce the best fit model using the Box-Jenkins method. We tend to get some good fits, but some weaknesses still exist in terms of the behaviour of the peaks in the ACF or PACF plots of the residuals. For Series 50 we got the chance to explore the seasonal ARIMA models as the time series showed non-stationary seasonality even after differencing.

In summary, we have covered all the aspects of descriptive analytics thoroughly and we have produced some good fitting models for each time series, ready to be used for forecasting!

1. **Table of Contents**

[**1.** **Executive Summary** 1](#_Toc119874679)

[**2.** **Table of Contents** 2](#_Toc119874680)

[**3.** **Introduction** 3](#_Toc119874681)

[**4.** **Series 22 (D449) Analysis** 3](#_Toc119874682)

[**4.1 Graphical Interpretation** 3](#_Toc119874683)

[**4.2 Statistical Test** 6](#_Toc119874684)

[**4.3 ACF & PACF Analysis** 7](#_Toc119874685)

[**5.** **Series 50 (D478) Analysis** 8](#_Toc119874686)

[**5.1 Graphical Interpretation** 8](#_Toc119874687)

[**5.2 Statistical Test** 11](#_Toc119874688)

[**5.3 ACF & PACF Analysis** 11](#_Toc119874689)

[**6.** **Series 78 (D506) Analysis** 14](#_Toc119874690)

[**6.1 Graphical Interpretation** 14](#_Toc119874691)

[**6.2 Statistical Test** 16](#_Toc119874692)

[**6.3 ACF & PACF Analysis** 17](#_Toc119874693)

[**7.** **Conclusion** 19](#_Toc119874694)

[**8.** **Appendices** 21](#_Toc119874695)

[**R Script** 21](#_Toc119874696)

[**Appendix 1** 24](#_Toc119874697)

[**Appendix 2** 25](#_Toc119874698)

[**Appendix 3** 26](#_Toc119874699)

[**Appendix 4** 26](#_Toc119874700)

[**Appendix 5** 27](#_Toc119874701)

[**Appendix 6** 27](#_Toc119874702)

[**Appendix 7** 28](#_Toc119874703)

1. **Introduction**

We were allocated with 3 different time series obtained from the M4 competition dataset. These time series originated from real-world businesses but were transformed to avoid any negative values and maintain the anonymity of the data used. The main objective of the assignment is to explore the data to its best without going into the forecasting stage. To this end, this report will only comprise of the descriptive analytics of the data.

The data exploration involves analysing data in its raw form i.e., daily values as the M4-info sheet implies. Due to the frequency being too large, we were unable to pinpoint a clear timeframe of when the changes in the time series were happening. To rectify this matter, we have then explored the data by aggregating it to three levels: weekly, monthly, and quarterly for easier visualisation. Additionally, we have also plotted a yearly time series but as this level only produced 10 or 11 observations for the time series, the possibility of capturing any seasonal patterns was more limited and thus, we have opted to remove this aggregate.

Generally, the analysis involves plotting the time series, checking for various regular components such as trend, seasonality, and level shift, identifying irregular components such as the presence of any outlier, decomposing the time series to get more insights of the data, performing statistical testing to check for stationarity of the data, and lastly, conducting ACF & PACF analysis to find the most suitable models for forecasting. The time series which will be covered in this report are Series 22 (D449), Series 50 (D478) and Series 78 (D506).

1. **Series 22 (D449) Analysis**

## **4.1 Graphical Interpretation**

Chart, histogram

Description automatically generated

Figure 4.1: Time Series Plot at Daily Aggregate Level of 3,581 days dated from 3/1/2000 until 22/10/2009.

From Figure 4.1 we can see that nothing can be commented about the trend and seasonal components, but we can witness a few outliers. We can visually see the presence of outliers, but the test of outliers is not in the scope of this report. We do not have any obvious reasons or opinions about the presence of these outliers since we do not know what the time series represents.

We will now aggregate the data at various levels and try to analyse the regular components in detail. The approach of obtaining the values of the different aggregate levels is by averaging the values according to their corresponding aggregate level. We are using mean as the function parameter for our aggregate levels in R. This will give us the mean value of that set of data at a time. We must not ignore the possibility of some variations in the mean value due to the presence of outliers but for now, we will proceed with mean values only. The different aggregate levels of the time series were then plotted and shown in the figure below:

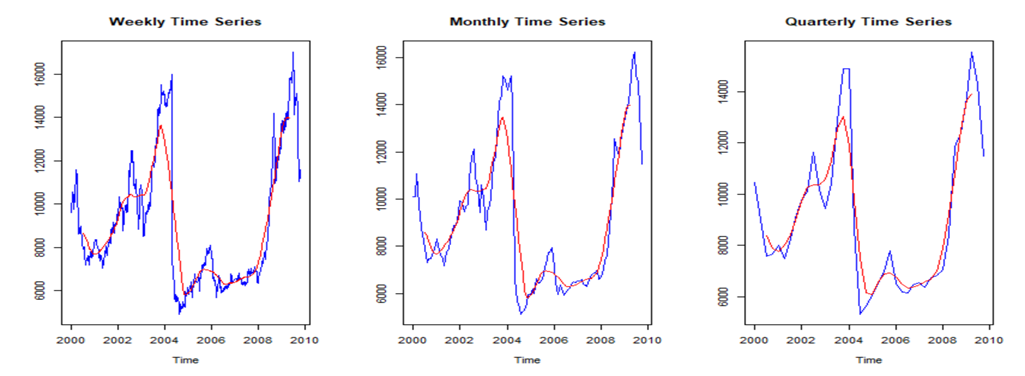


Figure 4.2: Weekly Time Series Plot with CMA of 52 in red (left), Monthly Time Series Plot with CMA of 12 in red (middle) and Quarterly with CMA of 4 in red (right).

We have plotted 3 time series of different aggregate levels along with their CMA to visually identify the trend of the time series. The red line in the graph is the trend without any seasonality. The trend pattern for all these time series is upward at the beginning, then a huge dip due to which the trend tends to be downward followed by a short constant pattern and then an upward pattern again at the end. Due to this dip and the presence of the outliers we can also conclude that there is a level shift in our time series.

Data from the year 2000 to 2004 is at a certain level and suddenly the huge drop in value just after the beginning of the year 2004 resulted in moving the entire data to a different level, which is certainly lower than the previous level. So, data from the year 2004 to 2008 is the result of a level shift and hence we can see a significant change in trend. After 2008, the trend again follows an upward pattern but again in 2010 there seems to be a dip, this could probably be due to the unavailability of the complete data for that year. We will now speculate more about the seasonality in our different time series.

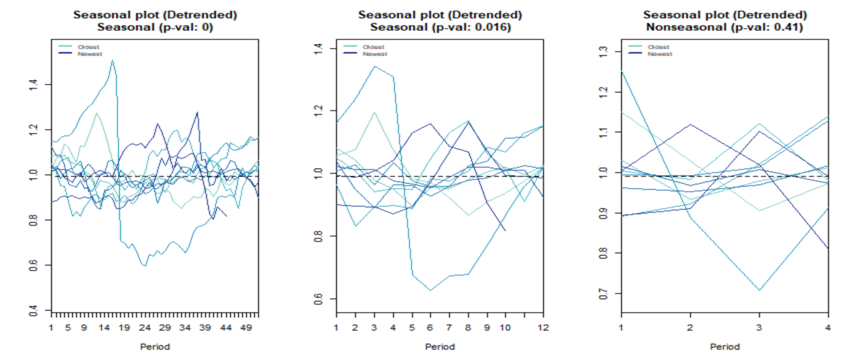


Figure 4.3: Seasonal Plots of Weekly (left), Monthly (middle) and Quarterly (right).

Based on Figure 4.3, the seasonal plot for the quarterly time series (right) shows no clear seasonality both from graphical observation and from the statistical aspect as the p-value of 0.41 is greater than the significance level value, thus failing to reject the null hypothesis of the time series being non-seasonal. However, for the monthly and weekly time series, the seasplot function implies that they are seasonal with a p-value of 0 and 0.016 respectively. The seasonal patterns in both time series are not clearly visible in the plots so we will further decompose these time series to get more details about the seasonal component.

We have chosen the classical decomposition approach to identify the patterns of the monthly time series using the decomposition function available in the “tsutils” package in R. As the time series is not explicit in showing whether it is an additive or multiplicative model, we have decomposed it using both types of models before determining which one is the best suited. The seasonal component of each decomposition is separately analysed. Along with patterns, the values in both decompositions are significant enough to prove that there exists some seasonality in the time series.

We will apply multiplicative and additive decomposition one by one on both the series i.e., weekly and monthly. If we separately observe the season component of both decompositions (multiplicative and additive) for the weekly and monthly series, we can observe that their decompositions show some seasonality but the values in weekly are comparatively more significant than the values in the monthly time series. In the case of the multiplicative decomposition of the weekly time series, the seasonal component shows approximately 4% of the seasonal effect whereas in the case of the multiplicative decomposition of the monthly time series the seasonal component shows approximately 2.5% of the seasonal effect.

Similarly, the seasonal component of additive decomposition of the weekly time series shows a rise of approximately 400 in the original data value whereas, for the monthly time series, the seasonal component shows a rise of approximately 200. So, we chose to go ahead with exploring the weekly time series. The multiplicative and additive decomposition of weekly time series is plotted below. Refer to Appendix 1 for the decomposition plots of the monthly time series.

Chart, line chart

Description automatically generated

Figure 4.4: Additive Decomposition Model for Weekly Time Series

Chart, line chart

Description automatically generated

Figure 4.5: Multiplicative Decomposition Model for Weekly Time Series

We couldn’t decide on a model just by visualising the decomposition plots here. Thus, we determined the model by comparing the error values resulting from the difference between the actual time series and the regular components (trend and seasonal). The results of the calculation are shown in the table below:

|  |  |  |
| --- | --- | --- |
| **Decomposition Type** | **Additive** | **Multiplicative** |
| Mean Error (ME) | -1.603e-14 | 69.578 |
| Mean of Absolute Error (MAE) | 761.995 | 752.915 |
| Mean of Absolute Percentage Error (MAPE) | 8.475% | 8.157% |

Table 1: Comparison of ME, MAE and MAPE values for both Additive and Multiplicative Decomposition

From Table 1, we can now identify Multiplicative as the best model as it has the lower value for all the MAE and MAPE.

## **4.2 Statistical Test**

The time series must be in a stationary state in order to find a good forecasting model for it. KPSS and ADF tests are commonly used for this purpose. After performing these statistical tests on our weekly time series, we know that the time series is not stationary because the p-value in KPSS is 0.02005 (which is less than the critical value of 5%) so we can reject the null hypothesis of data being stationary. Similarly, the p-value in the ADF test is 0.4947 which means we fail to reject the null hypothesis of non-stationarity at 5%. This was expected as the time series exhibits some seasonality. We can see the tsdisplay output for our non-stationary time series in Appendix 2. The test results at the initial level are shown as the following:

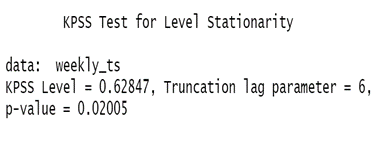
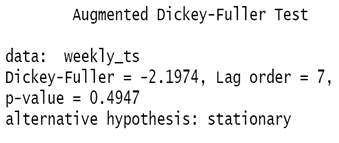
 

Figure 4.6: KPSS and ADF tests result for monthly time series

Since the time series is not stationary, we take the first differencing of the data and again perform KPSS and ADF tests on the differenced time series. Fortunately, after the first round of differencing only, we get to see that the time series is now stationary since the p-value in KPSS is greater than 0.1 and the p-value in ADF is smaller than 0.01.

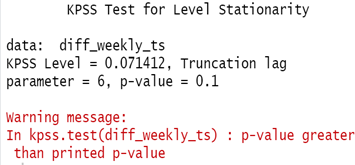
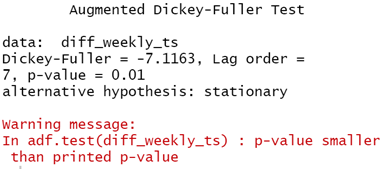
 

Figure 4.7: KPSS and ADF tests result for first level differencing

We are now good to go to analyse the ACF & PACF to identify the best fit models for forecasting our time series.

## **4.3 ACF & PACF Analysis**

In order to proceed with finding the best forecasting model for this weekly time series, we opted to use the Box-Jenkins method before comparing our findings with a model suggested by the auto ARIMA function from R. We have already checked for the stationarity of the time series in the previous section and obtained that the time series becomes stationary after first differencing. So, we will now plot the ACF & PACF of the differenced time series in order to determine the orders of AR & MA manually.

Diagram

Description automatically generated

Figure 4.8: Plot of stationarity for first level differencing of weekly time series (top), ACF plot (bottom left) and PACF plot (bottom right).

We can determine the order of AR & MA by visualising the peaks in ACF and PACF functions respectively. The peaks in the PACF plot tell us the order of AR and it is better to start experimenting with orders of AR first before including the orders of MA in our ARIMA model. Since there is only one peak in PACF so we will start with ARI (1,1) i.e., Arima (1,1,0).

After we apply the ARIMA (1,1,0) on our monthly time series and plot the residuals of our fit we get that there aren’t any more peaks in the ACF and PACF plots. This shows that this ARIMA (1,1,0) model is a good fit for our time series. Please see the figure below:

Diagram

Description automatically generated with medium confidence

Figure 4.9: Plot of first fit residuals (top), ACF plot (bottom left) and PACF plot (bottom right).

By manually applying other combinations for ARIMA order like ARIMA(0,1,1) and ARIMA(1,1,1) and plotting the ACF and PACF of the residuals of the fit we find no more peaks in ACF and PACF function plots which concludes that even these models are a good fit. Refer to Appendix 3 for plots of residuals of these fits.

Finally, we use auto.arima function to get the order for ARIMA and we get the same order as what we manually chose at the beginning i.e., ARIMA(1,10). By trying different sets of auto.arima such as defining “ic” parameter as “aic” and defining the test parameter as “adf” test we again derive the same ARIMA order which is ARIMA(1,1,0).

1. **Series 50 (D478) Analysis**

## **5.1 Graphical Interpretation**

Chart, line chart, scatter chart

Description automatically generated

Figure 5.1: Time Series Plot at Daily Aggregate Level of 3,320 days dated from 30/3/2001 until 1/5/2010

Figure 5.1 shows a plot of a daily time series where we can roughly observe a stationary pattern at the start followed by a non-stationary pattern up until the end of the series. We have then decided to start aggregating the data. The initial approach of obtaining the values of the different aggregate levels was by averaging the values according to their corresponding aggregate level. However, for this time series, this approach failed in detecting any seasonal components (refer to Appendix 4). Therefore, we have chosen to use the summation or the total values instead. The first 2 rows of data dated 30th – 31st March 2001 and the last row dated 1st May 2010 were removed in order to avoid the summation values being influenced or skewed. Hence, the following time series will start from 1st April 2001 until 30th April 2010, consisting only of 3,317 days. The different aggregate levels of the time series are then plotted and shown in the figure below:

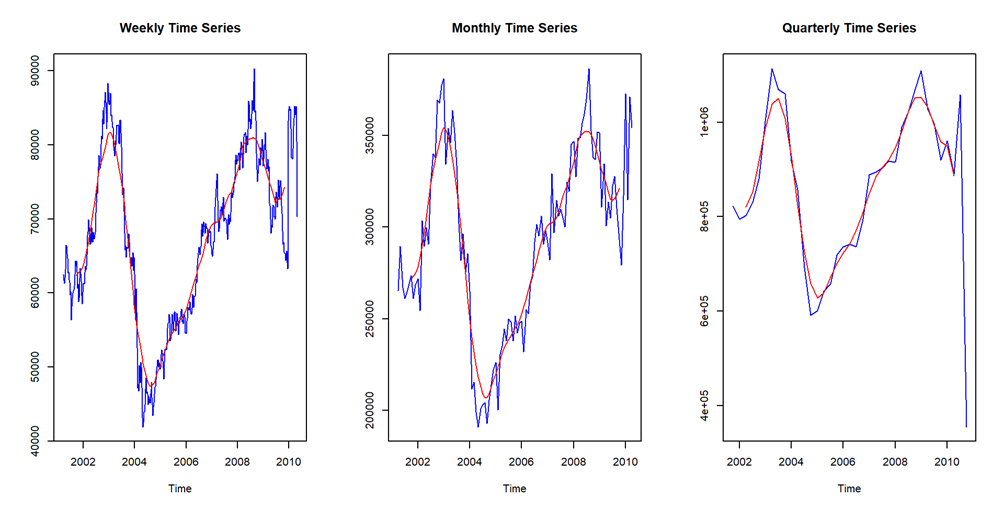


Figure 5.2: Weekly Time Series Plot with CMA of 52 in red (left), Monthly Time Series Plot with CMA of 12 in red (middle) and Quarterly Time Series Plot with CMA of 4 in red (right).

We have plotted 3 time series of different aggregate levels along with their CMA to visually identify the trend of the time series in Figure 5.2. The red line in the graph is the trend without any seasonality. The trend pattern for all these time series is upward at the beginning, followed by a downward pattern and later followed by an upward pattern again. At the end of the quarterly time series, we can see a huge dip which is highly due to the availability of only one month worth of data in the second quarter of the year 2010. We have then applied the seasplot function to all the aggregates to detect the seasonal component and the plots are shown in Figure 5.3.

Chart

Description automatically generated with medium confidence

Figure 5.3: Seasonal Plots of Weekly (left), Monthly (middle) and Quarterly (right) Time Series.

From Figure 5.3, the seasonal plot for the weekly time series shows no clear seasonality both from graphical observation and from the statistical aspect as the p-value of 0.997 is greater than the significance level value, thus failing to reject the null hypothesis of the time series being non-seasonal. This is also the case for the quarterly aggregate where no seasonality could be seen and proved from the p-value. However, for the monthly aggregate, we can roughly see from the plot that there is a seasonality albeit a weak one, where generally, there will be a peak at Month 1, followed by a dip in Month 2. There are also two other slightly lower dips that happen in Month 9 and 11, respectively. To further investigate the components of the monthly time series, a decomposition of the data has been conducted and the results are shown. From now onwards, we have proceeded with only the monthly time series for further exploration.

Chart, line chart

Description automatically generated

Figure 5.4: Additive Decomposition Model for Monthly Time Series 

Chart, line chart

Description automatically generated

Figure 5.5: Multiplicative Decomposition Model for Monthly Time Series

When solely comparing the graphs, it remains unclear which one is the most suitable model to decompose the monthly time series. The values are slightly significant for the additive seasonality ranging from approximately -30,000 to 10,000 for the highest peaks whereas the multiplicative model ranges from -0.7% to 1.03%. As both these ranges are relatively low and no conclusion can be made, the alternative way to determine the model is by comparing the error values resulting from the difference between the actual time series and the regular components (trend and seasonal). The results of the calculation are shown in the following table:

|  |  |  |
| --- | --- | --- |
| **Decomposition Type** | **Additive** | **Multiplicative** |
| Mean Error (ME) | -4.272e-12 | 597.545 |
| Mean of Absolute Error (MAE) | 9,839.969 | 9,947.816 |
| Mean of Absolute Percentage Error (MAPE) | 3.401% | 3.422% |

Table 2: Comparison of ME, MAE and MAPE values for both Additive and Multiplicative Decomposition.

From the table, we can identify Additive as the best model as it has the lower value for all the mean errors calculated.

## **5.2 Statistical Test**

KPSS and ADF tests here are conducted on the monthly time series as we have established that this time series can capture the seasonal component compared to the other aggregate levels. The first round of tests gives the same result that the time series is not stationary which was expected based on our previous findings. The results of the tests are as below:

Text

Description automatically generated with medium confidence  Text

Description automatically generated

Figure 5.6: KPSS and ADF tests result for monthly time series

After a first round of differencing, the results for the KPSS test and ADF test did not agree with each other, the former proved to be stationary with a p-value greater than 0.1 while the latter proved to be non-stationary with a p-value as 0.07301 as shown below:

Text

Description automatically generated Text

Description automatically generated

Figure 5.7: KPSS and ADF tests result for first differenced data

Therefore, another round of differencing is needed to counter this inconclusiveness. Based on the ACF plot of this differenced model, there were significant peaks in lags 12, 24 and 36 which suggested a non-stationary seasonality. Therefore, a seasonal differencing was carried out and once again was tested for stationarity. Unfortunately, only the KPSS test again proved the stationarity of this differenced time series (refer Appendix 5).

Due to this, we have then used the ndiffs and nsdiffs functions available in R to see if the results would be similar to our findings. The nsdiffs suggested only 1 differencing while the ndiffs suggested 0 seasonal differencing. Due to this, we have decided to follow the suggestion as it is also highly unadvisable to have high orders of differencing. Thus, the stationarity of the data will only be based on the KPSS test in order to proceed with the next stage of ACF and PACF analysis.

## **5.3 ACF & PACF Analysis**

We will first use the Box-Jenkins method to manually create the best fit model using the ARIMA function in R for the monthly time series. We now must analyse both the ACF and PACF plots to determine the order for the ARIMA model. Below is the tsdisplay output for the first differenced time series:

Chart

Description automatically generated with medium confidence

Figure 5.8: Plot of stationarity for first level differencing of monthly series (top), ACF plot (bottom left) and PACF plot (bottom right).

Based on Figure 5.8, we can observe the seasonal peaks at lags 12, 24 and 36 in the ACF plot for the differenced data as expected hence further supporting that a seasonal ARIMA model would be the best fit model for this time series. There are also some notable negative peaks in the ACF plot at lags 1, 11 and 23. Meanwhile, in the PACF plot, we could see 2 negative peaks at lags 1 and 11 which exceeds the significance bounds (dashed line in blue).

We started with a simple seasonal ARIMA model for our first fit before building up to a more complex model. The first fit model created considers the negative peak at lag 1 in the PACF plot so we will be testing an AR(1) model for both the non-seasonal and seasonal components therefore the model would be ARIMA(1,1,0)(1,0,0)[12] and the outputs are shown in Figure 5.9.

Diagram

Description automatically generated

Figure 5.9: Plot of ARIMA(1,1,0)(1,0,0)[12] residuals (top), ACF plot (bottom left) and PACF plot (bottom right).

For this model, we can observe that most of the peaks are within the bounds for both ACF and PACF plots. However, there is a significant negative peak at lag 2 for both plots. Therefore, we will now try fitting another seasonal ARIMA model for the second fit using a combination of AR(2) in the seasonal component and MA(2) in the non-seasonal component which is ARIMA(1,1,2)(2,0,0)[12]. Chart, box and whisker chart

Description automatically generated

Figure 5.10: Plot of ARIMA(1,1,2)(2,0,0)[12] residuals (top), ACF plot (bottom left) and PACF plot (bottom right).

Based on Figure 5.10, it seems that the second fit is better than the former as almost all the peaks are within bounds. The AIC value of this fit (2396.62) is also smaller compared to the first fit (2430.13). To confirm if this is the best fit model for the monthly time series, we have then used the auto ARIMA function available in R to compare the outcome. Below is the result of the auto ARIMA function:

Chart, box and whisker chart

Description automatically generated

Figure 5.11: Residual plot for autofit ARIMA model (top), ACF plot (bottom left) and PACF plot (bottom right).

The autofit model created by this function is ARIMA(0,1,0)(2,0,0)[12]. All the peaks in both ACF and PACF are within bounds except for lag 11 in the PACF. However, this could also just indicate a random error as there is still a 5% possibility of this occurring by chance. The AIC value of this model also turned out to be the lowest at only 2413.18. Furthermore, we have also tried this auto.arima function with the additional parameter of “ic” and “adf” which resulted in the same model being generated as the one above. Therefore, we have decided to choose this model as the best fit model for this monthly time series compared to the ones manually created using the Box-Jenkins method.

1. **Series 78 (D506) Analysis**

## **6.1 Graphical Interpretation**

Chart, line chart, histogram

Description automatically generated

Figure 6.1: Time Series Plot at Daily Aggregate Level of 3,320 days dated from 30/03/2001 until 01/05/2010.

Figure 6.1 plots the original daily time series. Since the initial data set is daily data for 10 consecutive years, its high frequency has made trend and seasonality visually hard to observe, so it needs to be further transformed into a lower frequency time series and therefore we are aggregating this time series to weekly, monthly, and quarterly levels.

Chart, histogram

Description automatically generated

Figure 6.2: Weekly Time Series Plot with CMA of 52 in red, Monthly Time Series Plot with CMA of 12 in red, and Quarterly with CMA of 4 in red.

We have plotted the time series for three aggregate levels along with their CMA which depicts the trendline. There are a few dips which probably gives us the sense of the presence of some outliers in our data. It can be observed from the plots of the time series and the trendline that there is a possible cyclical fluctuation roughly on a two-year base present in the period where the first one lasted for 2 years starting from 2002 until 2004 while the other lasted for approximately 4 years from 2006 until 2010. Due to the presence of upward and downward peaks there is also a possibility of level shifts in our time series at period 2004 until 2006. We will now move on to inspect if there is any seasonality present in our different time series.

Graphical user interface, chart, diagram, histogram

Description automatically generated

Figure 6.3: Seasonal Plots of Weekly (left), Monthly (middle) and Quarterly Time Series (right).

Based on Figure 6.3, the seasonal plot for the quarterly time series and monthly time series shows no clear seasonality both from the graphical observation and from the statistical aspect as the p-value 0.137 and 0.292 respectively are greater than the significance level value, thus failing to reject the null hypothesis of the time series being non-seasonal. However, the weekly time series shows some seasonality as the p-value is 0.023 i.e., lower than the significance level value, thus rejecting the null hypothesis.

Since we have chosen two seasonal time series for further exploration in the previous sections, we are now choosing a non-seasonal time series to do further exploration. We are now performing the decomposition of our monthly and quarterly time series to see if there is any underlying seasonality. After analysing the seasonal components of our monthly and quarterly decompositions, we can conclude that though decomposition tries to show some seasonal patterns, they aren’t significant enough. The seasonality from the weekly time series completely disappeared in the monthly and quarterly time series. It might indicate that the time series does not show pronounced seasonality or that it achieves greater regularity with less noise when irregular patterns are removed by a lower frequency. Thus, we have decided to go with monthly time series for the rest of the data exploration.

Also, in our opinion, the frequency of monthly time series is greater than the frequency of quarterly time series, which will give us a clearer picture of the regular components. We can also clearly witness the presence of outliers in such cases. The multiplicative and additive decompositions of monthly time series are plotted below. We are trying to identify the most suitable model for our time series.

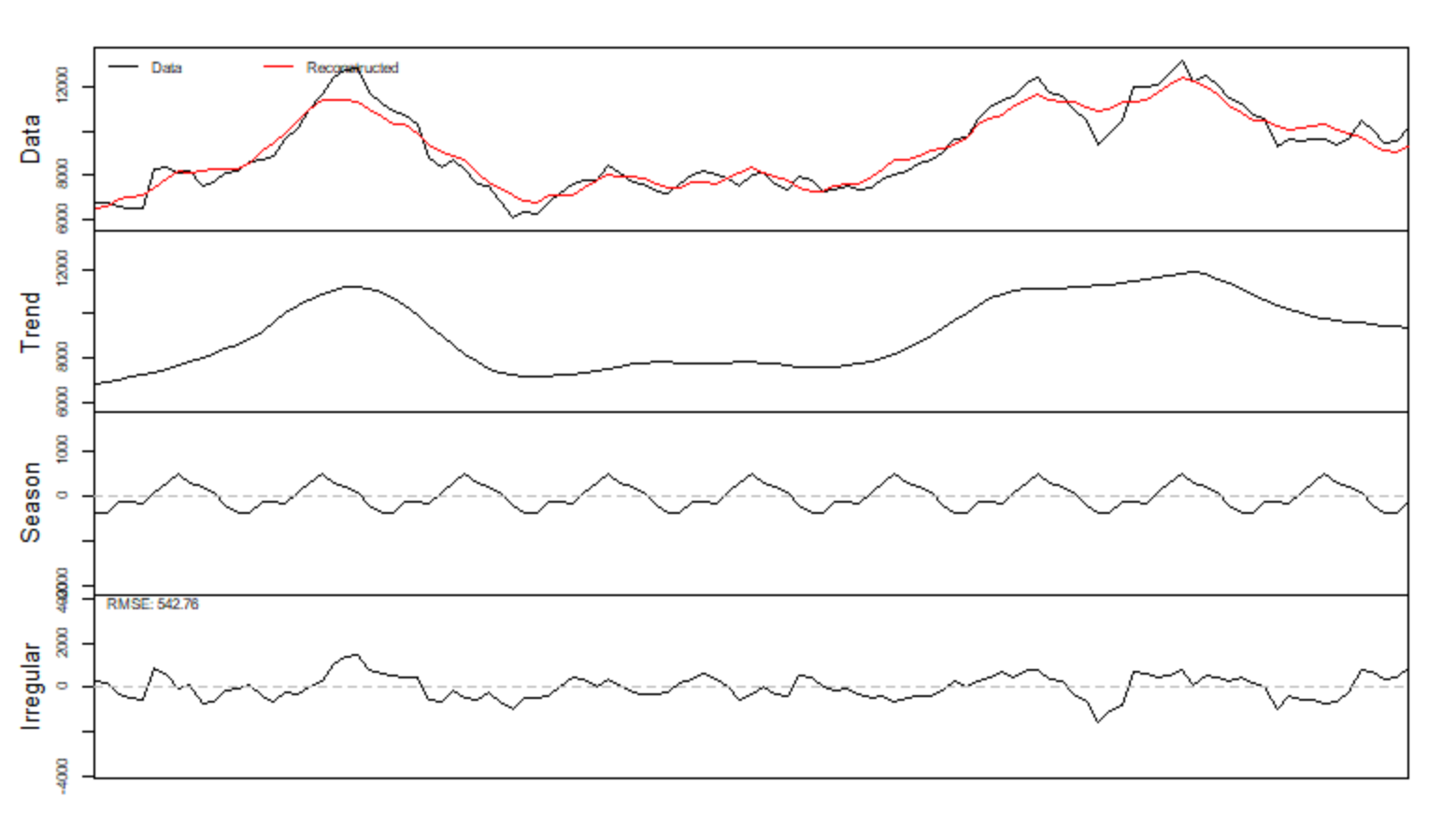


Figure 6.4: Additive Decomposition Model for Monthly Time Series

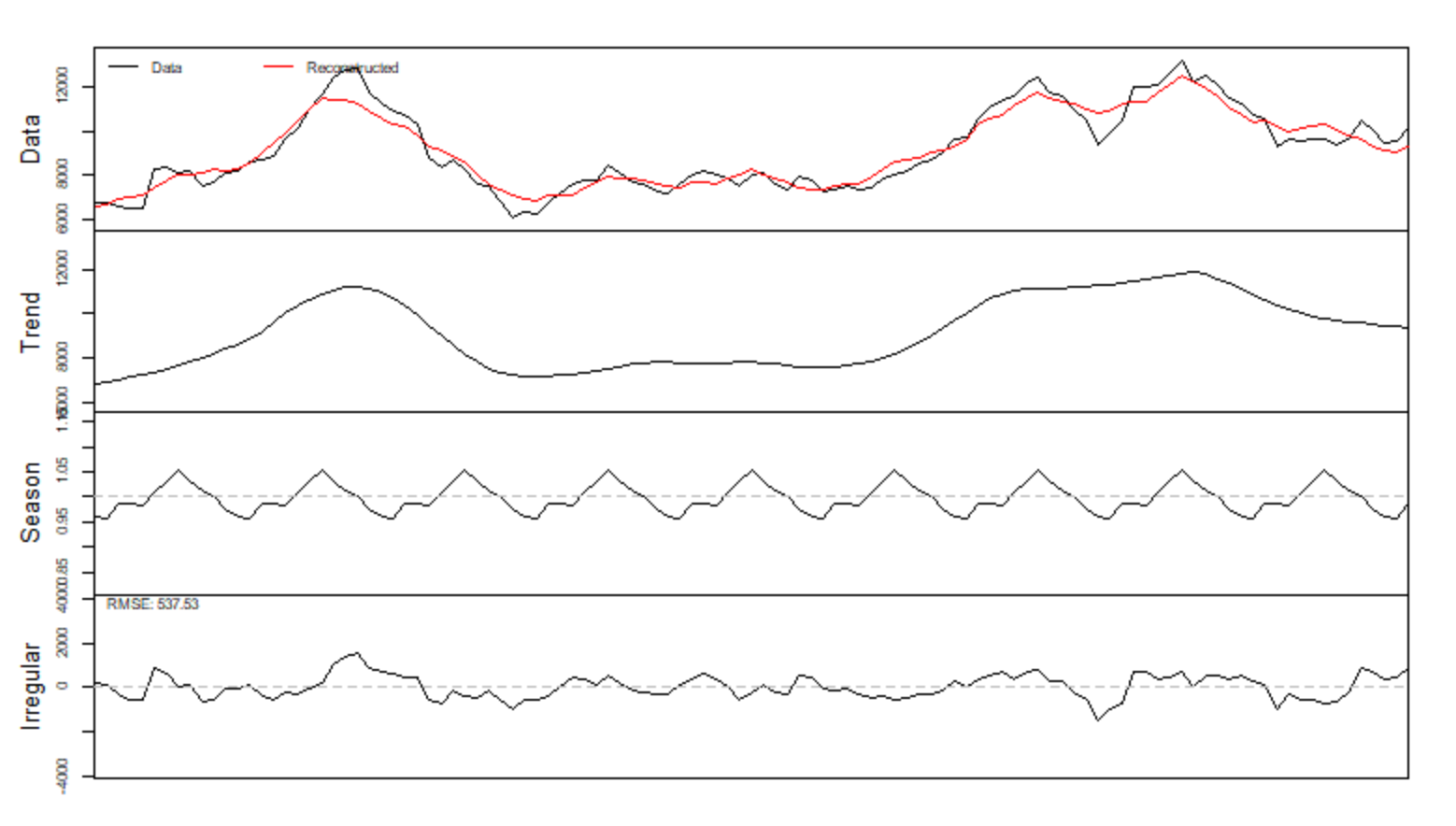


Figure 6.5: Multiplicative Decomposition Model for Monthly Time Series

By seeing the decomposition plots we can’t decide which is the most suitable model for our time series. We are thus going with another way of calculating mean error values. The results of the calculation are shown below:

|  |  |  |
| --- | --- | --- |
| **Decomposition Type** | **Additive** | **Multiplicative** |
| Mean Error (ME) | -8.184e-15 | 25.960 |
| Mean of Absolute Error (MAE) | 454.615 | 447.729 |
| Mean of Absolute Percentage Error (MAPE) | 4.979% | 4.880% |

Table 3: Comparison of ME, MAE and MAPE values for both Additive and Multiplicative Decomposition.

From the table, we can identify multiplicative as the better model as it has the lower value for MAE and MAPE values.

## **6.2 Statistical Test**

After performing the statistical tests i.e., KPSS and ADF on our monthly time series we discovered that the KPSS test proves that the time series is stationary. However, the p-value in ADF is 0.1777 (Figure 6.6) which proves that data is non-stationary and thus we have contradictory results from KPSS and ADF tests. This can be seen in the tsdisplay output for our non-stationary time series in the Appendix 6. The figure below shows the result of both the stationarity test of the series at the initial level:

**Graphical user interface

Description automatically generated with medium confidence** **Text

Description automatically generated**

Figure 6.6: KPSS and ADF tests result for monthly time series

Since the time series is not stationary, we took the first differencing of the data and again performed the KPSS and ADF tests on the differenced time series. Fortunately, both tests proved that the differenced data is now stationary since the p-value in KPSS is larger than 0.1 and the p-value in ADF is smaller than 0.01. (Figure 6.7)

Graphical user interface, text

Description automatically generated Graphical user interface, text, application

Description automatically generated

Figure 6.7: KPSS and ADF tests result for first differenced time series

We could now move on to the ACF & PACF analysis for the identification of best fit models for forecasting.

## **6.3 ACF & PACF Analysis**

As per the previous times series, we will first carry out the Box Jenkins method to determine the best forecasting model for our time series. Using the stationary first level of differenced data, we have plotted the ACF and PACF (Figure 6.8) of our differenced time series to determine the order of AR and MA for our ARIMA model.

Chart, box and whisker chart

Description automatically generated

Figure 6.8: Plot of stationarity for first level differencing of monthly time series (top), ACF plot (bottom left) and PACF plot (bottom right).

We can determine the order of AR & MA by visualising the significant spikes in the ACF and PACF plots, respectively. Since the PACF plot has one peak at the very beginning of the lag, we started with an ARI(1,1) model i.e. ARIMA(1,1,0) and plot the ACF and PACF of the residuals as follows:

Chart, box and whisker chart

Description automatically generated  
Figure 6.9: Plot of ARIMA(1,1,0) residuals (top), ACF plot (bottom left) and PACF plot (bottom right).

From Figure 6.9, there are two other apparent peaks in the PACF plot corresponding to lags 18 and 19. These are greater than the frequency of the months, therefore it does not imply that seasonal ARIMA should be used instead, which also aligns with the findings where no evidence of seasonality exists in the monthly time series. However, due to the presence of this unexplainable peaks, the first fit model is not highly satisfactory, and we decided to create another model of ARIMA(1,1,1) to see if this would be a better fit.

Chart, box and whisker chart

Description automatically generated

Figure 6.10: Plot of ARIMA(1,1,1) residuals (top), ACF plot (bottom left) and PACF plot (bottom right).

Unfortunately, the residual ACF and PACF plots of the second fit shows little to no difference from the first fit model where the peaks in lags 18 and 19 still appeared. By manually applying other combinations for ARIMA order such as ARIMA(0,1,1) and ARIMA(2,1,0) and plotting the ACF and PACF of the residuals of these fits, we could still hardly see any changes in comparison to the residuals plots shown in Figure 6.9 and Figure 6.10 (Refer to Appendix 7 for these plots of other ARIMA models). Therefore, this could mean that the peaks observed in the residual plots are just random errors which exist in the differenced monthly time series that prevents the ARIMA model being tested from fitting well.

Finally, we used the auto.arima function to cross-check on the ARIMA order suggested and it returned an ARIMA(1,1,0) model, which is the same as our first fit. By trying different sets of auto.arima parameters such as defining “ic” as “aic” and defining test parameter as “adf”, we again derived the same ARIMA order which is ARIMA(1,1,0).

Hence, even though the suggested ARIMA model has its weakness which is the presence of random errors but it still produced the lowest AIC value, so we would conclude that ARIMA(1,1,0) should be used for forecasting.

1. **Conclusion**

In general, all the time series analysed in this report displayed a non-stationary property throughout the whole period of either 3,581 days for Series 22 or 3,320 days for both Series 50 and 78. This is expected, as real-world time series tend to experience upward and downward trends over 9 or 10 years. As the data is anonymous, we could not relate or find concrete reasons behind any of the trends or outliers for our time series. There were also no restrictions in terms of having a set forecast horizon, therefore we were free to explore and choose the best aggregate level to use for the forecast model.

For Series 22, it could be stated that a clear level shift happened in the time series due to the presence of outliers (huge dip in time series plot) whereas for Series 78 level shift had happened in the early half instead. Series 50 seemed to have a constant level at the beginning of the time series but then it followed an upward trend reaching its first peak. All the time series present an increasing and decreasing trend throughout the whole period except for the part where it remains stationary as stated above. When doing the aggregation, both Series 22 and 78 used the mean value apart from Series 50 where the total value was used instead. Since the summation value was used, the last quarter for this time series seemed like an outlier, however, this was only due to the availability of only 1 month of data instead of 3.

A seasonal component was present in the Weekly aggregate for both Series 22 and 78, in the Monthly Aggregate for both Series 22 and 50, and none in the Quarterly aggregate for all the time series. From this discovery, we have then decided to use the Weekly aggregate for Series 22 and the Monthly aggregate for Series 50 which contain seasonality and use the Monthly aggregate for Series 78 which lacks this component to see if this would affect the outcome of the forecast model created. Further data decomposition using the decomp function in R has aided us in determining whether our time series were either an additive or multiplicative model. From this function itself, it was still unclear, as the values of both the additive and multiplicative season components were relatively low, therefore we have calculated the ME, MAE and MAPE for each of the models. We were then able to determine the model of each time series which is Additive for Series 50 and Mixed Multiplicative for both remaining time series.

Next, we used both KPSS and ADF tests to determine the stationarity of the time series. As expected, all the time series failed to prove that they were stationary. Fortunately, after only the first level of differencing, Series 22 and 78 achieved stationarity, whilst for Series 50, the results of the KPSS test proved stationarity, the ADF test proved otherwise. However, after using the nsdiffs and ndiffs function in R to check for the suggested level of differencing, the outcome was 1 and 0, respectively. With this, we decided to take up the suggestion and proceed with only basing the stationarity of the time series using the KPSS test. From the residual plots of each differenced data of the time series, only Series 50 displayed significant peaks in lags 12, 24 and 36 showing that a non-stationary seasonality exists and therefore a seasonal ARIMA would be more suited as a forecast model.

In the final stage of the exploration, we used the Box-Jenkins method to create some best fit models for the time series to be used for forecasting. For Series 22, we have managed to find the best fit model for the series which is ARIMA(1,1,0). This model was also the same one as suggested bythe auto.arima function available in R and it is deemed to be the best fit due to the nature of the peaks in both the ACF and PACF plots which were all within the significance bound.

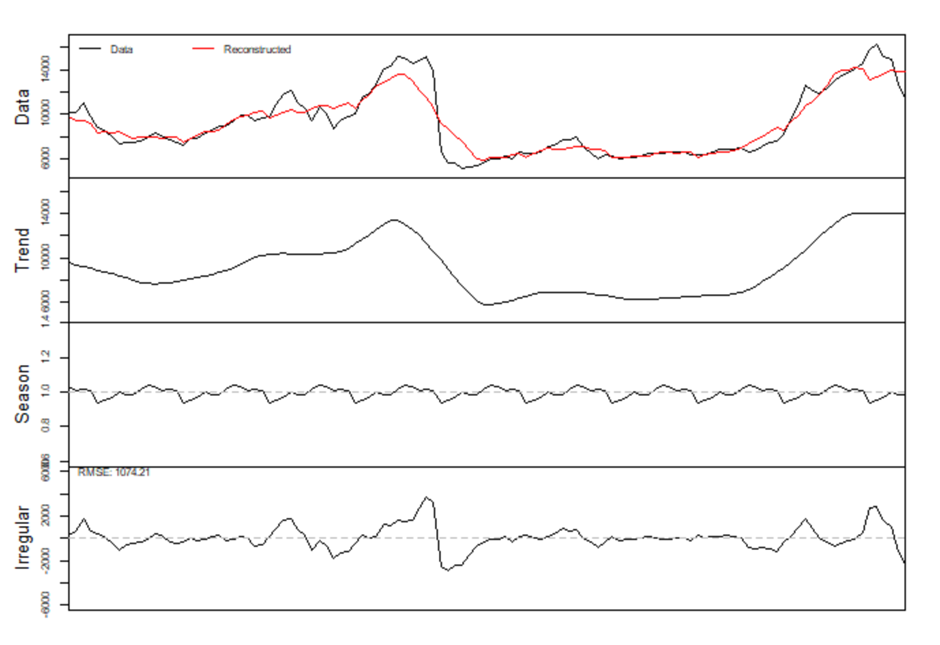
However, this was not the case for both Series 50 and 78. We created a first fit model of ARIMA(1,1,0)(1,0,0)[12] where there was still a significant peak in the ACF and PACF residual plots at lag 2. From this, we have created another model fit of ARIMA(1,1,2)(2,0,0)[12]. This was a better fit as the AIC value was lower and the peaks were all within the bounds. When we tried the auto.arima function to see if it will produce the same model, we were surprised when it gave a totall different model of ARIMA(0,1,0)(2,0,0)[12] but since the AIC value was the lowest compared to the others, we decided to conclude that this latter model would be a better fit for the time series.

For Series 78, we have created multiple models to fit the first differenced monthly time series ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(0,1,1) and ARIMA(2,1,0) where all of the residuals plots of both ACF and PACF still displayed high peaks at lag 18 and 19. The auto.arima function gave us an output of ARIMA(1,1,0) which was the same as our first fit model. This tells us that even with the “best” fit model, we could still observe some unusually significant peaks in the ACF and PACF plots which could imply that there will still be a possibility of random errors occurring.

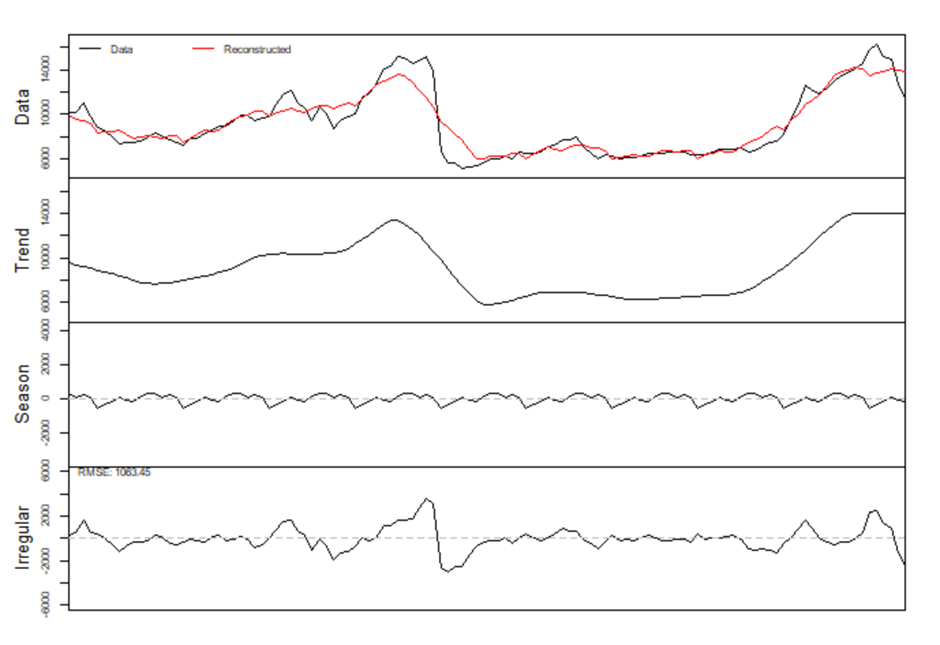
Lastly, when using a time series with seasonality, it is possible to have the “best” fit forecast model be either ARIMA or seasonal ARIMA and the “best” fit model might contain random errors buthave the lowest AIC.

1. **Appendices**

## **Appendix 1**

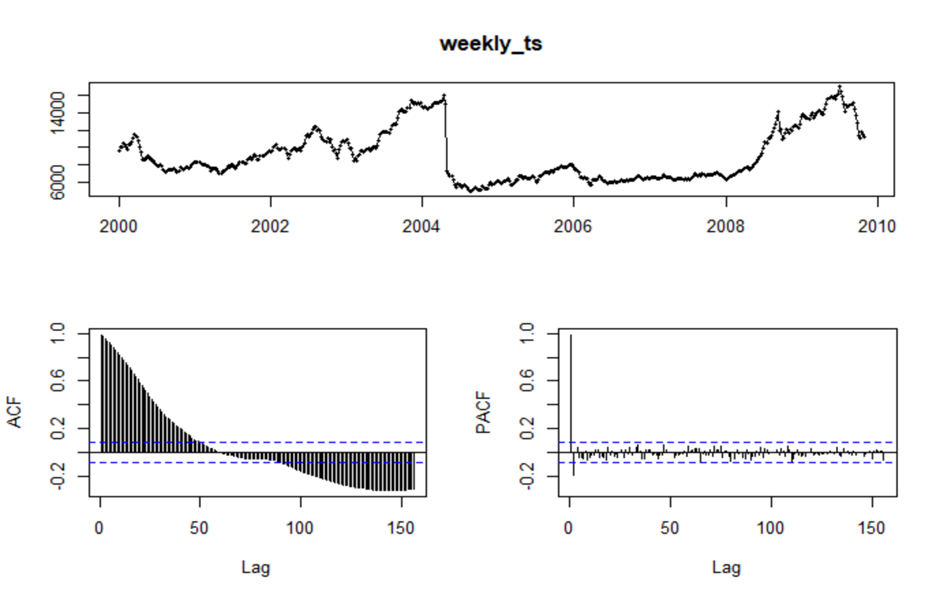


Multiplicative Decomposition Model for Monthly Time Series (22)



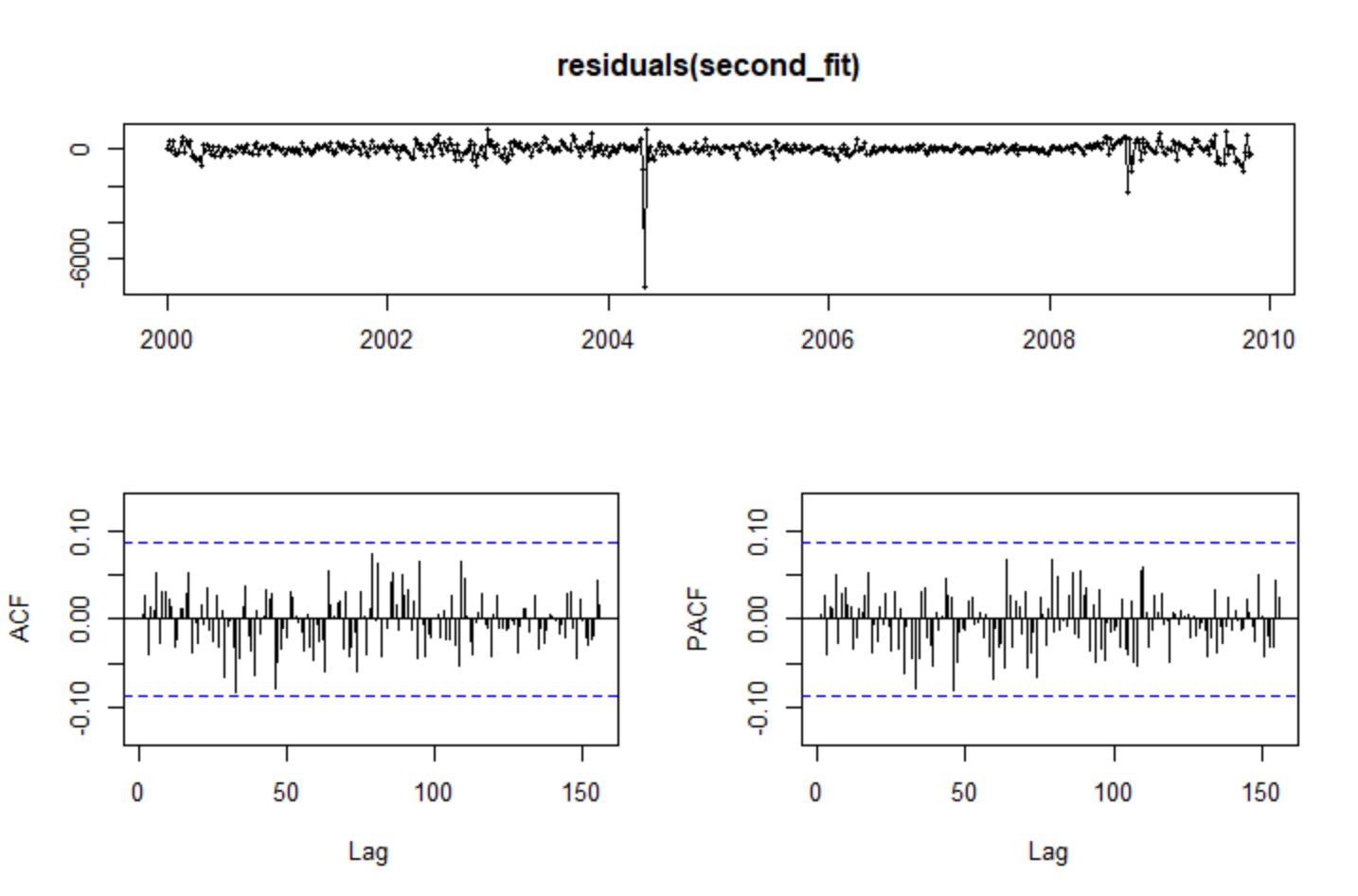
Additive Decomposition Model for Monthly Time Series (22)

## **Appendix 2**

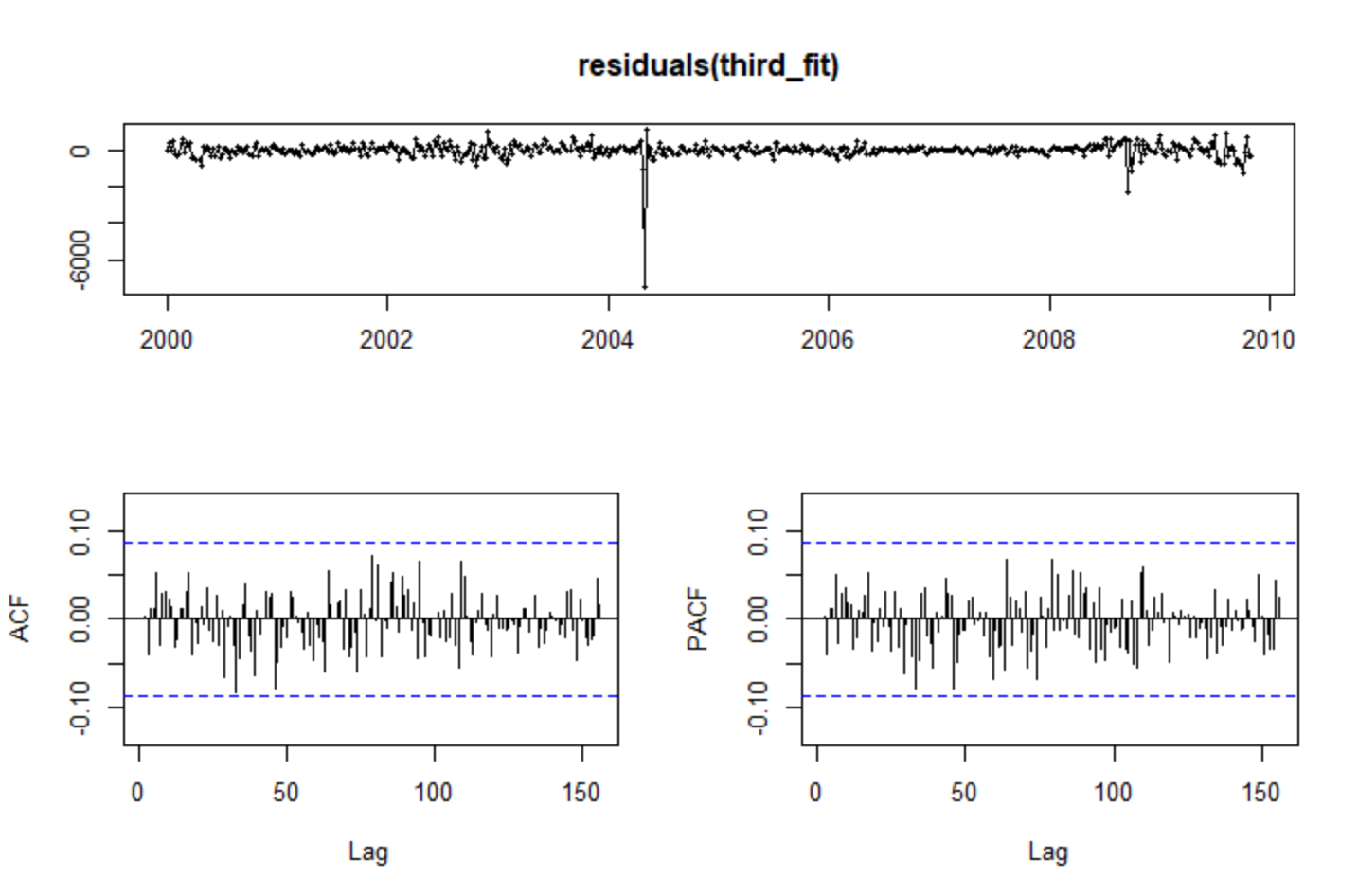


tsdisplay plot of weekly\_ts (series 22)

## **Appendix 3**



tsdisplay plot for ARIMA(0,1,1) of weekly time series of series 22



tsdisplay plot for ARIMA(1,1,1) of weekly time series of series 22

## **Appendix 4**

Graphical user interface, chart

Description automatically generated

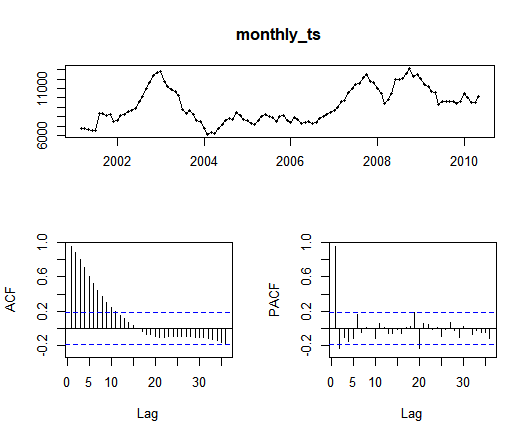
Seasonal plots for weekly, monthly and quarterly time series using the mean values for Series 50

## **Appendix 5**



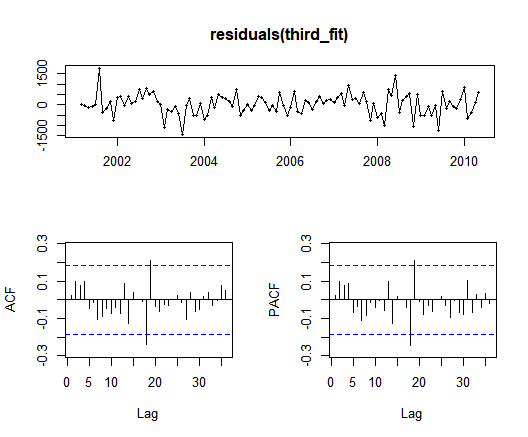
KPSS and ADF tests result for first level and seasonal differencing of monthly time series (50)

## **Appendix 6**

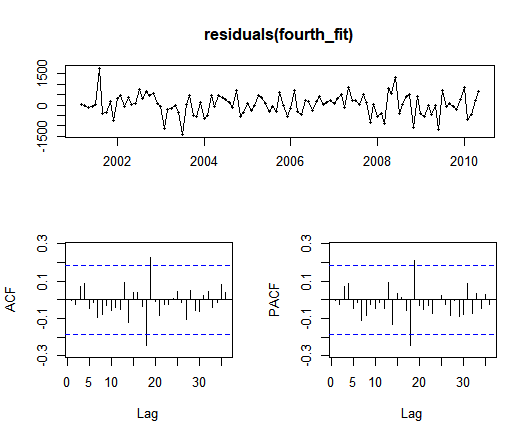


tsdisplay plot of monthly\_ts (series 78)

## **Appendix 7**



tsdisplay plot for ARIMA(0,1,1) of monthly time series of series 78



tsdisplay plot for ARIMA(2,1,0) of monthly time series of series 78