AI623 Time Series Analysis Group Project
Anti-diabetic Drug Sales Data-set

Teng Guang Way
G2102434F
teng0141@e.ntu.edu.sg

Hung Pun Ben
G2203078K
hung0017@e.ntu.edu.sg

Sean Goh Ann Ray
G2202190G
sean0057@e.ntu.edu.sg

Sim Ke Ying
G2203616G
S220106@e.ntu.edu.sg

Goh Jun Hao
G2102327H
gohj0091@e.ntu.edu.sg

1 PROJECT STATEMENT

The goal of this project is to showcase the ability to fit a suitable ARIMA model to real-world time series data. The expectations for this project include providing the background of the time series, explaining the process of developing an initial model by examining the ACF, PACF, differencing, transformations, and discussing how the model is improved by considering ACFs and PACFs in the residuals and employing ARIMA models. Additionally, the project will outline the steps taken to arrive at the final model and provide a forecast of the time series for three months or three years into the future.

2 DATASET

The data that the team decided on is the monthly anti-diabetic drug sales in Australia from 1992 to 2008. Total monthly scripts for pharmaceutical products falling under Anatomical Therapeutic Chemical (ATC) code A10, as recorded by the Australian Health Insurance Commission.
The following report will be focused on building a suitable model for the data-set and also to utilize that model to forecast the next three years ahead for the time series.

3 TIME PLOT OF ORIGINAL DATA-SET

Figure 1 below shows the original time plot of the original data-set. It can be observed to have a upward trending component in addition to a seasonal pattern which has increasing variance over time.

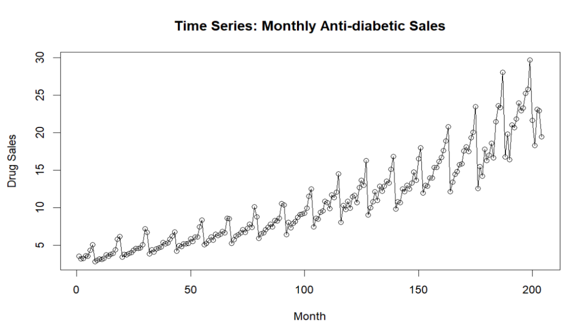

Figure 1: Time plot of original data

4 DATA TRANSFORMATION

4.1 Box-Cox Transformation

In order to stabilise the variance of the data, Box-Cox transformation is applied. From the BoxCox.lambda() function, lambda value of -0.0086589421 was derived. The lambda value used is close to zero lambda, which mean pure logarithmic function is applied to the data-set. It can be observed from the plot in Figure 2 that the variance is now stationary, however the trend and seasonal components still exist, which suggests further difference of data series is still required.

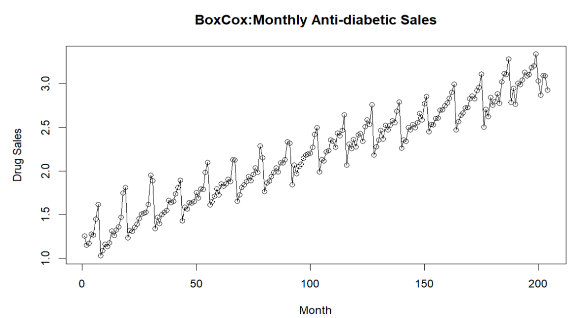

Figure 2: Time plot after Box-Cox transformation

4.2 Seasonal Differencing with Period 12

Seasonal differencing of period 12 is an operation used to remove seasonality and trend from a time series by taking the difference between the data points and their corresponding values 12 periods prior. This method is particularly useful for analyzing time series data with annual seasonality, such as monthly data with consistent patterns that repeat every year.

We applied seasonal differencing with period 12 on the box-cox transformation data. The difference data time plot visualization is as shown in Figure 3. The SACF are as shown in Figure 4. Both visualization techniques implies that the trend component and seasonal component are successfully removed by the lag 12 seasonal differencing. To further verify this, we use the formal hypothesis test to check for the presence of unit root and that data is stationary. The two functions adf.test() and kpss.test() were applied on the difference data. The returned p-value of ADF Test is 0.02. The returned p-value of KPSS Test is 0.1. For ADF-test, p-value of 0.02 suggest us to reject null hypothesis which states that unit root is present in the data, implying non-stationary. As such ADF-test also suggests that the series is stationary. For KPSS-test, p-value of 0.1 suggest to us that we fail to reject null hypothesis which states that series is trend-stationary.

This literally mean that both test further suggest to us that the difference data is stationary as per both unit root test.

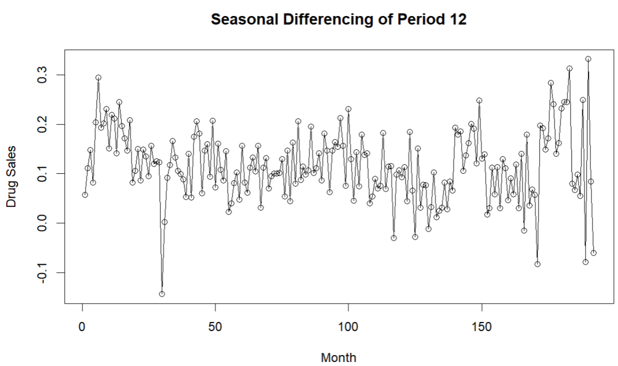

Figure 3: Time plot after seasonal difference with trend component and seasonal component removed

With the data now stationary, we will move on to plot the ACF and PACF of the transformed data to analyse which models to implement.

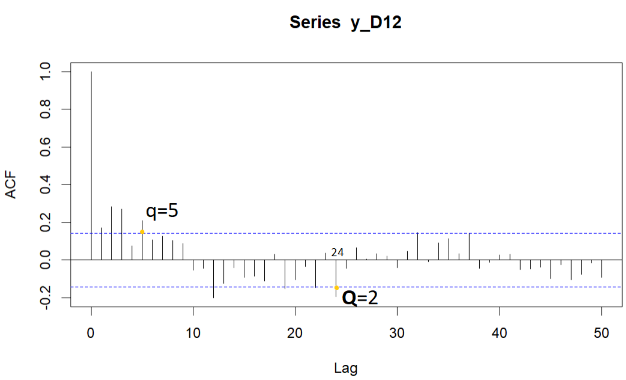

Figure 4: SACF Plot of the difference data

From the SACF plot in Figure 4, it can be observed that Q, which are observed lags at multiples of the seasonal period of 12, has a cut-off value of 2. On the other hand, q, which are observed lags starting from 0 up to the seasonal period of 12 minus 1, has a cut-off value of 5.

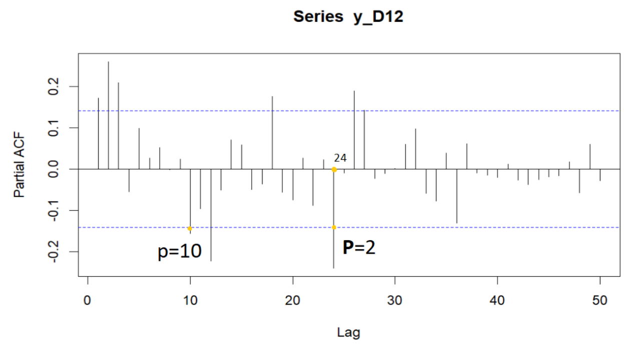

Figure 5: SPACF Plot of the difference data

Likewise, from the SPACF plot in Figure 5, it can be observed that P, which are observed lags at multiples of the seasonal period of 12, has a cut-off value of 2. On the other hand, p, which are observed lags starting from 0 up to the seasonal period of 12 minus 1, has a cut-off value of 10.

In summary based on the above findings we can conclude the following possible sarima arguments: p=10, q=5, O=2, Q=2.

5 SARIMA MODELS

Based on the p=10, q=5, P=2 and Q=2 values derived from the SACF and SPACF plots, 4 different SARIMA models fit1, fit2, fit3, fit4 were proposed and illustrated in section 5.1-5.4. Along the way, we also derived sub-models from some of those 4 models through further inspection of the residual SACF/SPACF.

5.1 Fit1: SARIMA((p=0, d=0, q=5, P=0, D=1, Q=2)

It can be observed that Fit1 model is not an adequate model as p-values of the Ljung-Box statistics are below 0.05 as shown in Figure 6. However we can still improve on the model by evaluating the residuals of Fit1. The SACF and SPACF of Fit1's residuals are shown in Figures 7 and 8 respectively. Looking at the SACF the cut off within the first 12-1 lags occurs at Lag 1 and the cut off in multiples of 12 lags occurs at Lag 2, implying q=1 and Q=2. Looking at the SPACF, the cut off at the first 12-1 lags occurs at Lag 4, and the cut off in multiples of 12 lags occurs at Lag 2, implying p=4 and P=2.

Therefore, with the new possible sarima arguments of p=4, q=1, P=2, Q=2 from the residuals SACF and SPACF, for Fit1 we can propose 4 new improvement models as:

1. sarima(p=0, d=0, q=6, P=0, D=1, Q=4),
2. sarima(p=4, d=0, q=5, P=2, D=1, Q=2),
3. sarima(p=0, d=0, q=6, P=2, D=1, Q=2),
4. sarima(p=4, d=0, q=5, P=0, D=1, Q=4).

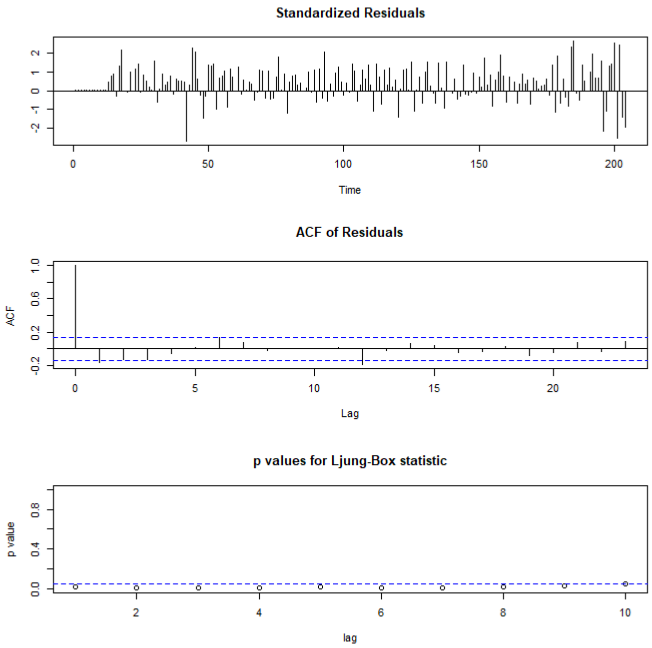

Figure 6: Fit1 sarima(p=0, d=0, q=5, p=0, D=1, Q=2) tsdiag result

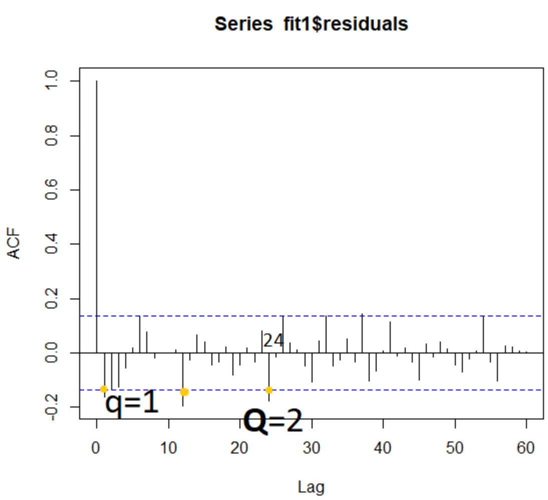

Figure 7: SACF Plot of Fit1 sarima(p=0, d=0, q=5, p=0, D=1, Q=2) residuals

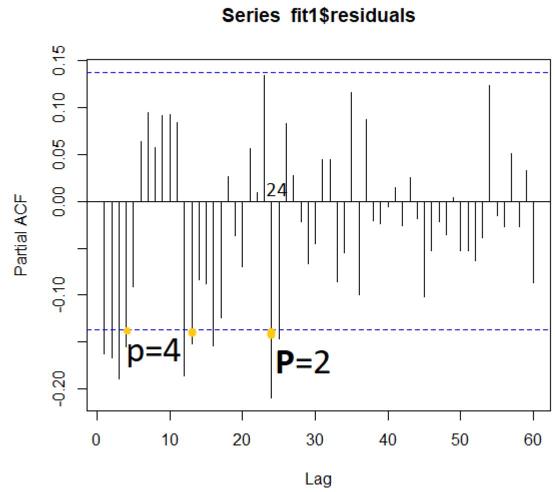

Figure 8: SPACF Plot of Fit1 sarima(p=0, d=0, q=5, p=0, D=1, Q=2) residuals

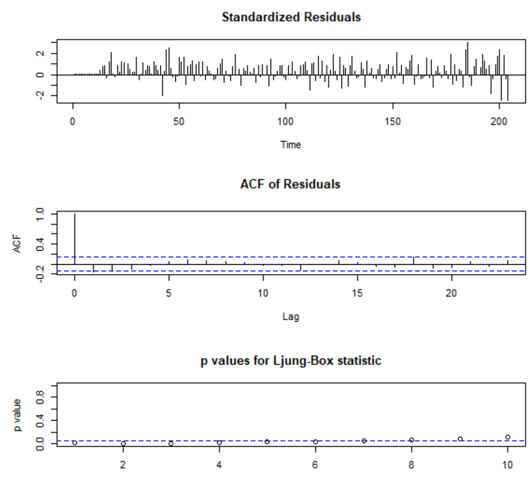

Figure 9: Fit1 Improvement 1 sarima(p=0, d=0, q=6, P=0, D=1, Q=4) tsdiag result

5.1.1 *Fit1's Improvement 1: sarima(p=0, d=0, q=6, P=0, D=1, Q=4).*

Applying the first improvement, sarima(p=0, d=0, q=6, P=0, D=1, Q=4), does not make the model adequate as results of tsdiag as shown in Figure 9. We can note that the p-value of the Ljung-Box Statistic Test is lesser than 0.05 for some lags. We continue to analyze the residuals SACF and SPACF as shown in Figure 10 and Figure 11. Looking at the SACF the cut off within the first 12-1 lags occurs at Lag 2 and the cut off in multiples of 12 lags occurs at Lag 0, implying q=2 and Q=0. Looking at the SPACF, the cut off at the first 12-1 lags occurs at Lag 9, and the cut off in multiples of 12 lags occurs at Lag 0, implying p=9 and P=0.

Similarly, by evaluating the cut off before Lag 12 - 1 and at multiples of 12, we can come up with a new set of possible sarima arguments where p=9, q=2, P=0 and Q=0. We can again further derive another 2 competing models:

1. sarima(p=0, d=0, q=8, P=0, D=1, Q=4)
2. sarima(p=9, d=0, q=6, P=0, D=1, Q=4)

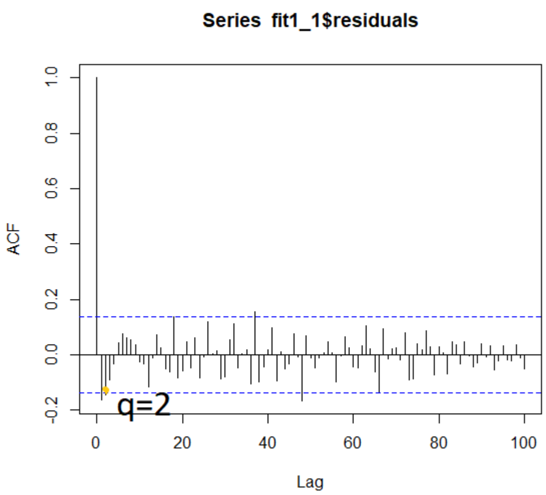

Figure 10: SACF Plot of Fit1 Improvement 1 sarima(p=0, d=0, q=6, P=0, D=1, Q=4) residuals

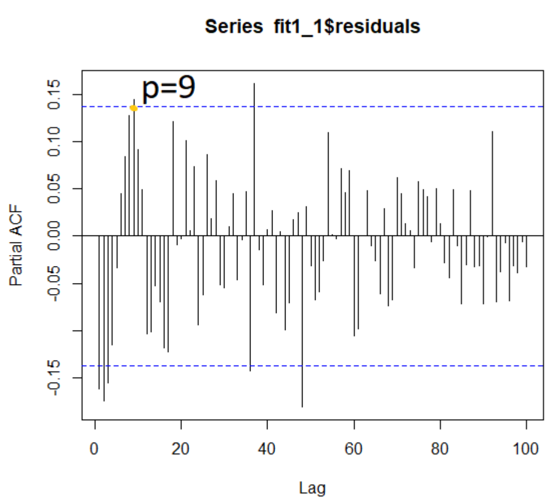

Figure 11: SPACF Plot of Fit1 Improvement 1 sarima(p=0, d=0, q=6, P=0, D=1, Q=4) residuals

We begin to attempt the further improvement on the failed improvement model using sarima(p=0, d=0, q=8, P=0, D=1, Q=4) and the tsdiag is as shown in Figure 12.

Again, after applying the further improvement, unfortunately the fitted model does not pass adequate test of Ljung-Box Statistic Test as per Figure 12 as p-values at some lags are below 0.05. The residuals SACF and SPACF again evaluated as shown in Figure 13 and Figure 14 respectively. Looking at the SACF the cut off within the first 12-1 lags occurs at Lag 9 and the cut off in multiples of 12 lags occurs at Lag 0, implying q=9 and Q=0. Looking at the SPACF, the cut off at the first 12-1 lags occurs at Lag 9, and the

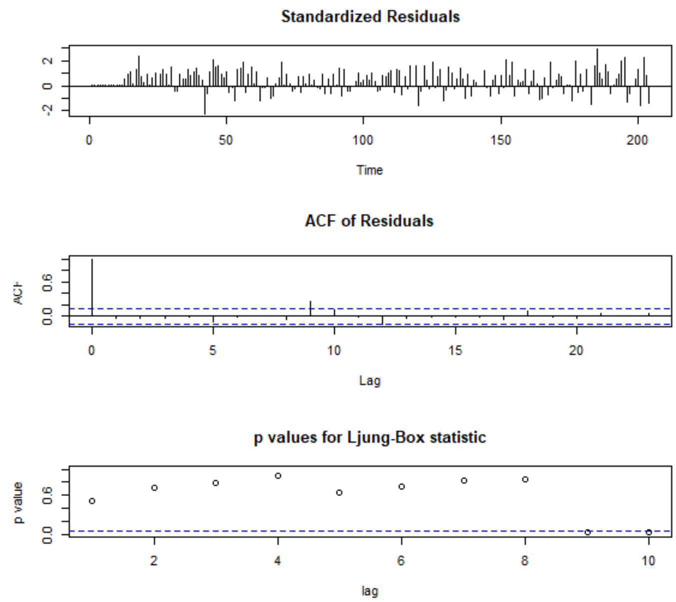

Figure 12: Fit1 Improvement 1 Further Improvement 1 sarima(p=0, d=0, q=8, P=0, D=1, Q=4) tsdiag result

cut off in multiples of 12 lags occurs at Lag 4, implying p=9 and P=4.

Again similarly with new set of possible sarima arguments where p=9, q=9, Q=0, P=4, we can derive another 4 competing models:

1. sarima(p=0, d=0, q=17, P=4, D=1, Q=4)
2. sarima(p=9, d=0, q=8, P=4, D=1, Q=4)
3. sarima(p=0, d=0, q=17, P=0, D=1, Q=4)
4. sarima(p=9, d=0, q=8, P=0, D=1, Q=4)

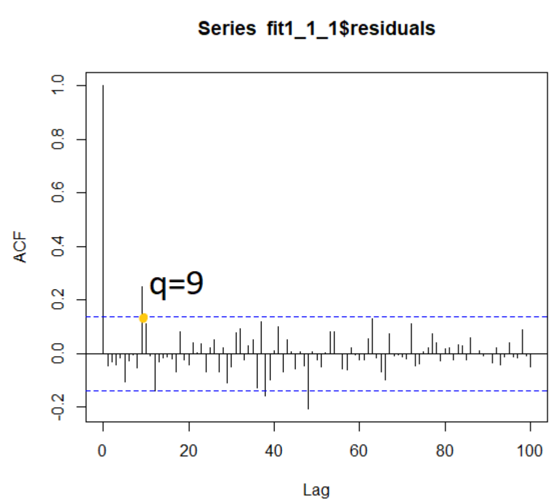

Figure 13: SACF Plot of Fit1 Improvement 1 Further Improvement 1 sarima(p=0, d=0, q=8, P=0, D=1, Q=4) residuals

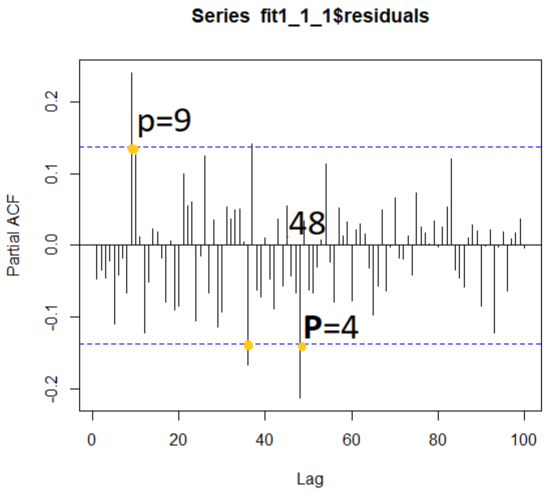

Figure 14: SPACF Plot of Fit1 Improvement 1 Further Improvement 1 sarima(p=0, d=0, q=8, P=0, D=1, Q=4) residuals

We begin to attempt the even further improvement on the failed improvement model using sarima(p=0, d=0, q=17, P=4, D=1, Q=4) and the tsdiag is as shown in Figure 15. Finally the sarima model passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals distribution is White Noise. The model has an AIC score of -499.0.

Then, we continue to attempt the other even further improvement on the failed improvement model using sarima(p=9, d=0, q=8, P=4, D=1, Q=4) and the tsdiag is as shown in Figure 16. This tells us that the fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. The model has an AIC score of -511.74.

Then, we continue to attempt the other even further improvement on the failed improvement model using sarima(p=0, d=0, q=17, P=0, D=1, Q=4) and the tsdiag is as shown in Figure 17. This tells us that the fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. The model has an AIC score of -490.55.

Then, we continue to attempt the other even further improvement on the failed improvement model using sarima(p=9, d=0, q=8, P=0, D=1, Q=4) and the tsdiag is as shown in Figure 18. This tells us that the fitting passes the adequate check as per the Ljung-Box Statistic

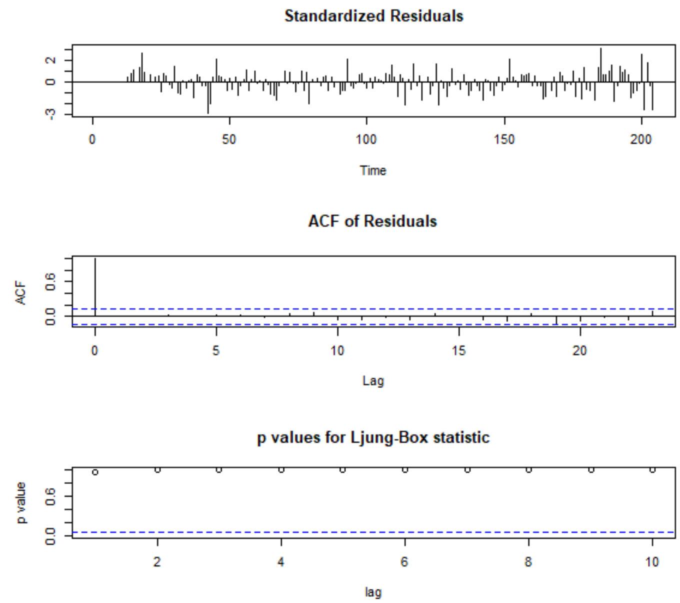

Figure 15: tsdiag of Fit1 Improvement 1 Further Improvement 1 Even Further Improvement 1, tsdiag result for sarima(p=0, d=0, q=17, P=4, D=1, Q=4)

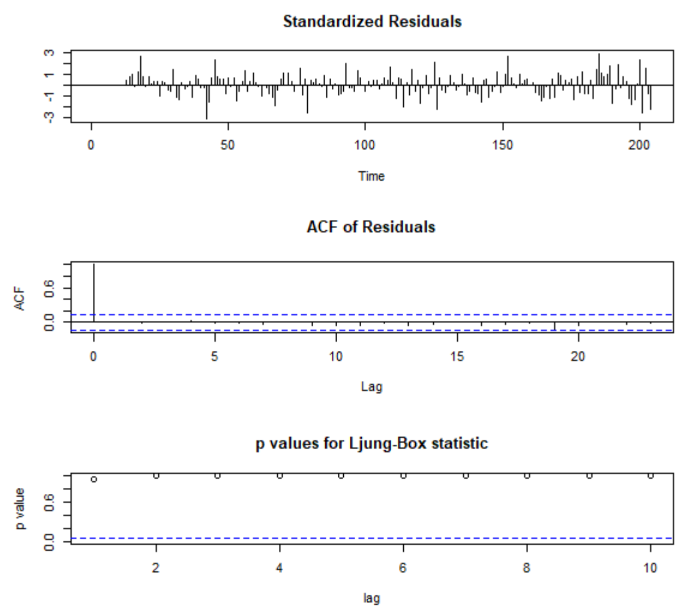

Figure 16: Fit1 Improvement 1 Further Improvement 1 Even Further Improvement 2, sarima(p=9, d=0, q=8, P=4, D=1, Q=4) tsdiag result

Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF

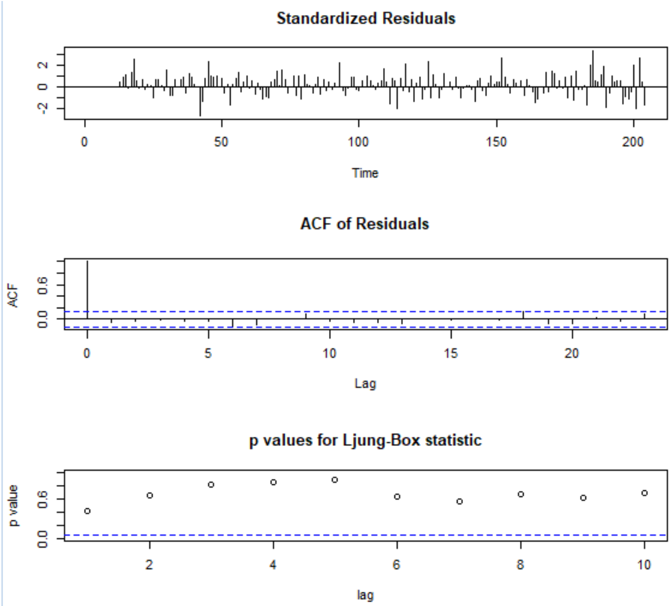

Figure 17: Fit1 Improvement 1 Further Improvement 1 Even Further Improvement 3, sarima(p=0, d=0, q=17, P=0, D=1, Q=4) tsdiag result

cut off instantly, implying that the residuals are White Noise. The model has an AIC score of -518.11.

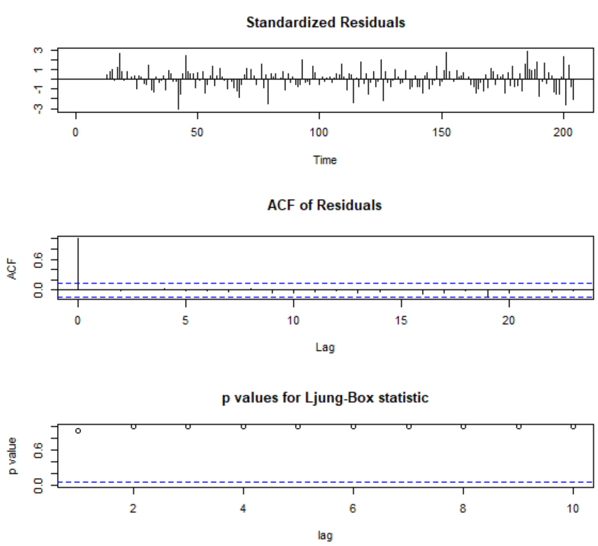

Figure 18: Fit1 Improvement 1 Further Improvement 1 Even Further Improvement1, sarima(p=9, d=0, q=8, P=0, D=1, Q=4) tsdiag result

Then, we continue to attempt the other further improvement on the failed improvement model using sarima(p=9, d=0, q=6, P=0, D=1, Q=4) and the tsdiag is as shown in Figure 19. This tells us that the fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. The model has an AIC score of -521.36.

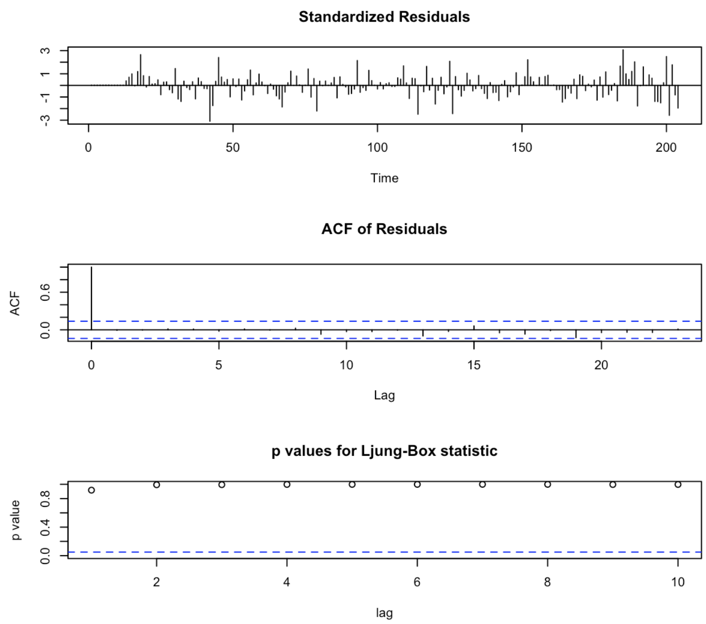

Figure 19: Fit1's Improvement 1 Further Improvement 2 sarima(p=9, d=0, q=6, P=0, D=1, Q=4) tsdiag result

For Fit1's Improvement 1, in summary we derived 5 valid model fittings.

1. sarima(p=0, d=0, q=17, P=4, D=1, Q=4) with AIC score -499
2. sarima(p=9, d=0, q=8, P=4, D=1, Q=4) with AIC score -511.74
3. sarima(p=0, d=0, q=17, P=0, D=1, Q=4) with AIC score -490.55
4. sarima(p=9, d=0, q=8, P=0, D=1, Q=4) with AIC score -518.11
5. sarima(p=9, d=0, q=6, P=0, D=1, Q=4) with AIC score -521.36

5.1.2 Fit1's Improvement 2: sarima(p=4, d=0, q=5, P=2, D=1, Q=2). After applying the second improvement sarima the model is adequate as results of tsdiag as shown in Figure 20. The fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. This model return us an AIC score of -519.37.

5.1.3 Fit1's Improvement 3 sarima(p=0, d=0, q=6, P=2, D=1, Q=2). Applying the third improvement does not make the model fitting adequate. Result of tsdiag, Ljung-Box Statistic test as shown in Figure 21 as the p-value at some lags is below 0.05. We continue to analyze the SACF and SPACF of residuals as shown in Figure 22 and Figure 23. Looking at the SACF the cut off within the first 12-1 lags occurs at Lag 9 and the cut off in multiples of 12 lags occurs at

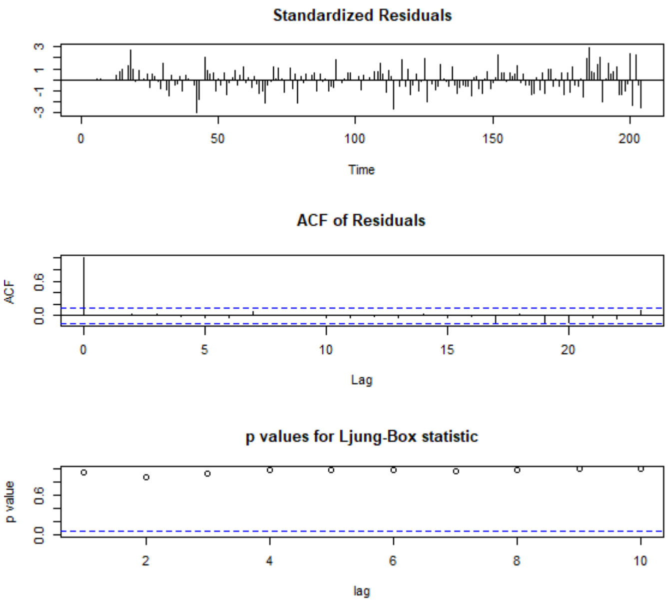

Figure 20: Fit1 Improvement 2 sarima(p=4, d=0 ,q=5, P=2, D=1, Q=2) tsdiag result

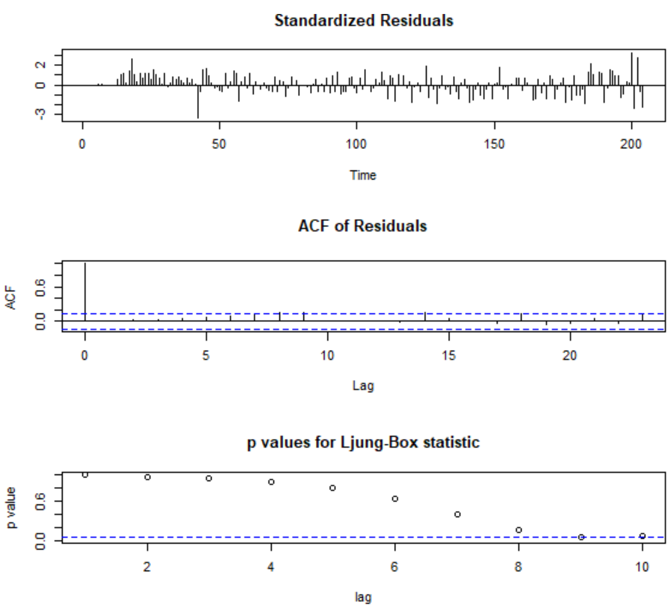

Figure 21: Fit1's Improvement 3 sarima(p=0, d=0, q=6, P=2, D=1, Q=2) tsdiag result

Lag 0, implying q=9 and Q=0. Looking at the SPACF, the cut off at the first 12-1 lags occurs at Lag 9, and the cut off in multiples of 12 lags occurs at Lag 0, implying p=9 and P=0.

Again similarly, based on the cut-off of the SACF and SPACF, we retrieve a new set of possible sarima arguments where p=9, q=9, Q=0, P=0 to derive another 2 competing models:

1. sarima(p=0, d=0, q=15, P=2, D=1, Q=2)
2. sarima(p=9, d=0, q=6, P=2, D=1, Q=2)

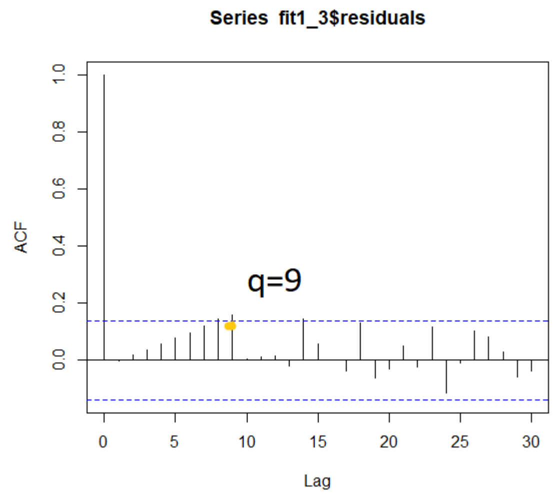

Figure 22: SACF Plot of Fit1 Improvement 3 sarima(p=0, d=0, q=6, P=2, D=1, Q=2)'s residuals

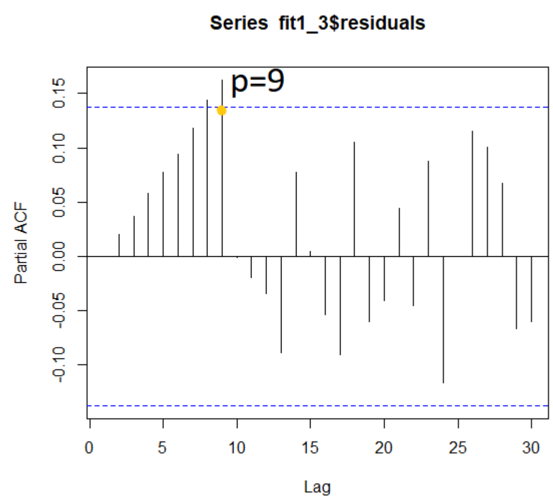

Figure 23: SPACF Plot of Fit1 Improvement 3 sarima(p=0, d=0, q=6, P=2, D=1, Q=2)'s residuals

Then, we begin to attempt the further improvement on the failed improvement model using sarima(p=0, d=0, q=15, P=2, D=1, Q=2). The tsdiag output is shown as per Figure 24. The fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. This model return us an AIC score of -484.51.

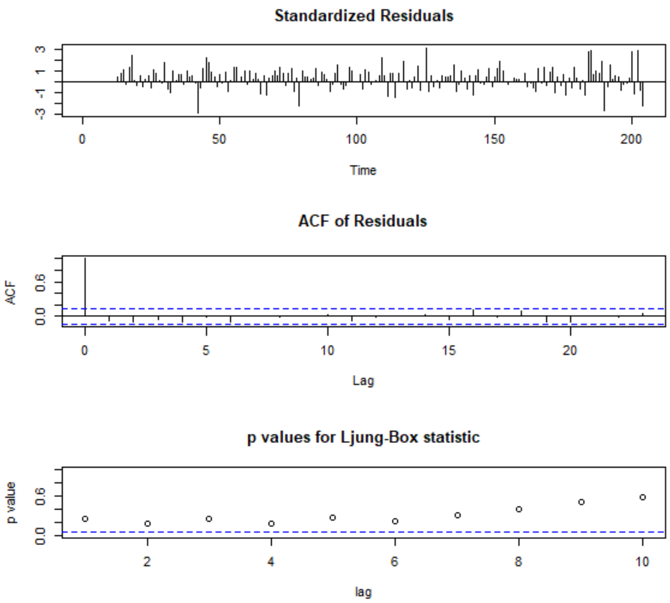

Figure 24: Fit1 Improvement 3 Further Improvement 1 sarima(p=0, d=0, q=15, P=2, D=1, Q=2) tsdiag result

Then, we continue to attempt the further improvement on the failed improvement model using sarima(p=9, d=0, q=6, P=2, D=1, Q=2). As per the tsdiag output, the series pass adequate test as shown in Figure 25. The fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. This model return us an AIC score of -520.96.

For Fit1's Improvement 3, in summary we derived 2 valid model fittings.

1. sarima(p=0, d=0, q=15, P=2, D=1, Q=2) with AIC score -484.51
2. sarima(p=9, d=0, q=6, P=2, D=1, Q=2) with AIC score -520.96

5.1.4 Fit1's Improvement 4: sarima(p=4, d=0, q=5, P=0, D=1, Q=4). Applying the forth improvement, sarima(p=4, d=0, q=5, P=0, D=1, Q=4). The model fitting passes adequate test as per the tsdiag return as shown in Figure 26. The fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. This model return us an AIC score of -520.

5.1.5 Fit1 SARIMA((p=0, d=0, q=5, P=0, D=1, Q=2) Conclusion. Even though the original Fit1, SARIMA((p=0, d=0, q=5, P=0, D=1, Q=2) is not a valid fitting, we managed to come up with 9 valid improved fittings which are adequate and pass the tsdiag diagnostic check. The fittings as followed:

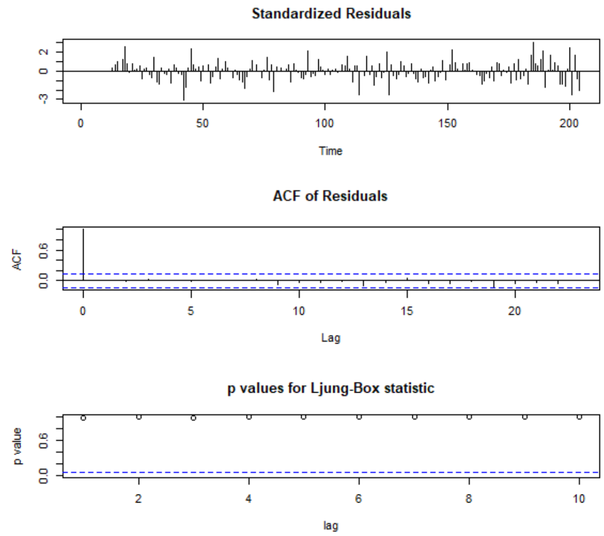

Figure 25: Fit1 Improvement 3 Further Improvement 2 sarima(p=9, d=0, q=6, P=2, D=1, Q=2) tsdiag result

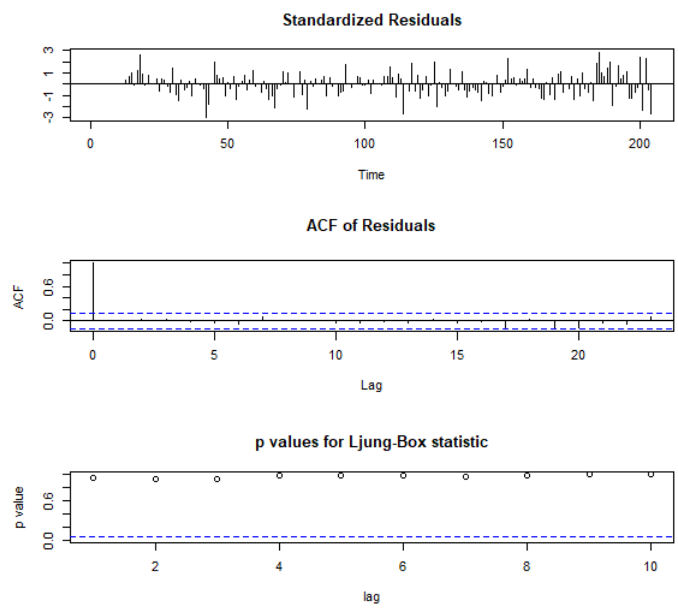

Figure 26: Fit1 Improvement 4 sarima(p=4, d=0, q=5, P=0, D=1, Q=4) tsdiag result.

1. sarima(p=0, d=0, q=17, P=4, D=1, Q=4) with AIC score -499
2. sarima(p=9, d=0, q=8, P=4, D=1, Q=4) with AIC score -511.74
3. sarima(p=0, d=0, q=17, P=0, D=1, Q=4) with AIC score -490.55
4. sarima(p=9, d=0, q=8, P=0, D=1, Q=4) with AIC score -518.11
5. sarima(p=9, d=0, q=6, P=0, D=1, Q=4) with AIC score -521.36
6. sarima(p=4, d=0, q=5, P=2, D=1, Q=2) with AIC score -519.37
7. sarima(p=0, d=0, q=15, P=2, D=1, Q=2) with AIC score -484.51

8. sarima(p=9, d=0, q=6, P=2, D=1, Q=2) with AIC score -520.96
9. sarima(p=4, d=0, q=5, P=0, D=1, Q=4) with AIC score -520

5.2 Fit2: SARIMA((p=0, d=0, q=5, P=2, D=1, Q=0)

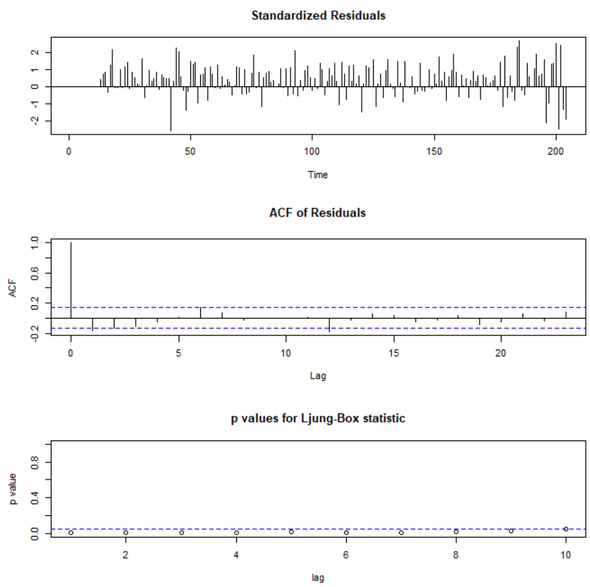

Figure 27: Fit2 sarima(p=0, d=0, q=5, P=2, D=1, Q=0) tsdiag result

From the above plot, it can be observed that Fit2 model is also not adequate model as p-values of the Ljung-Box statistics are below 0.05 as shown in Figure 27. However, we can still improve on the model by evaluating the residuals of Fit2. The SACF and SPACF of Fit2's residuals are shown in Figures 28 and 29 respectively. Looking at the SACF the cut off within the first 12-1 lags occur at Lag 1 and the cut off in multiples of 12 lags occur at Lag 2, implying q=1 and Q=2. Looking at the SPACF, the cut off at the first 12-1 lags occur at Lag 4, and the cut off in multiples of 12 lags occur at Lag 2, implying p=4 and P=2.

Therefore, for Fit2 we can propose 4 new improvement models with the new set of possible sarima arguments, p=4, q=1, P=2, Q=2 as:

1. sarima(p=0, d=0, q=6, P=2, D=1, Q=2)
2. sarima(p=4, d=0, q=5, P=4, D=1, Q=0)
3. sarima(p=0, d=0, q=6, P=4, D=1, Q=0)
4. sarima(p=4, d=0, q=5, P=2, D=1, Q=2)

5.2.1 Fit2's Improvement 1: sarima(p=0, d=0, q=6, P=2, D=1, Q=2). Applying the first improvement sarima(p=0, d=0, q=6, P=2, D=1, Q=2). The model fitting still fails the adequate test. The results produced of the Ljung-Box Statistic test is as shown in Figure 30. We continue to analyze the residuals SACF and SPACF as shown in Figure 31 and Figure 32. Looking at the SACF the cut off within the first 12-1 lags occurs at Lag 9 and the cut off in multiples of 12 lags occurs at Lag 0, implying q=9 and Q=0. Looking at the

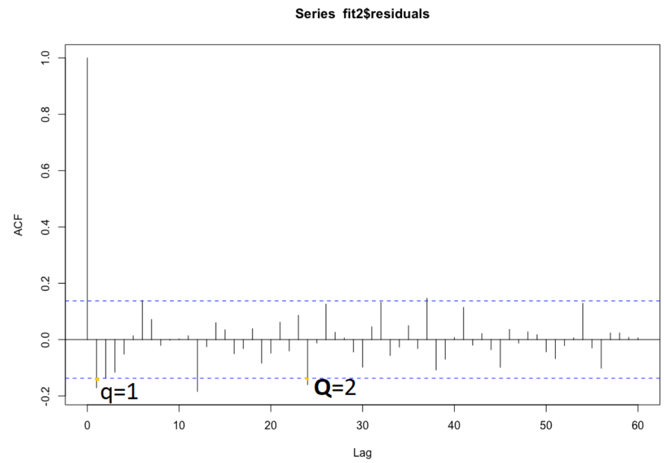

Figure 28: SACF Plot of Fit2 sarima(p=0, d=0, q=5, P=2, D=1, Q=0) residuals

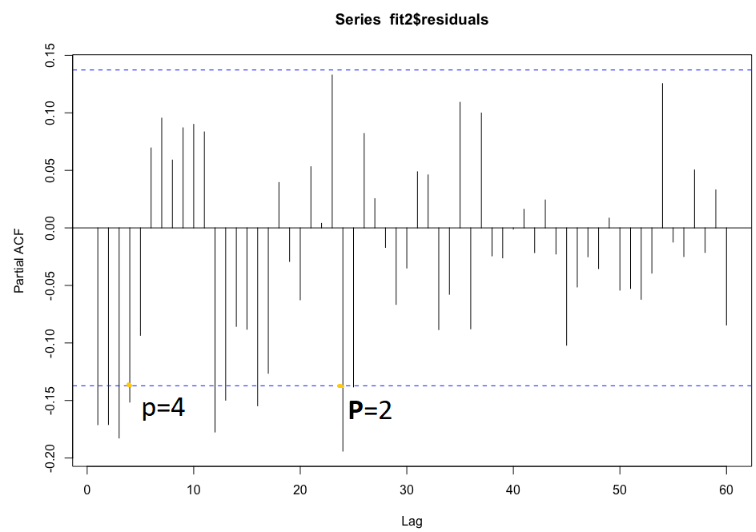

Figure 29: SPACF Plot of Fit2 sarima(p=0, d=0, q=5, P=2, D=1, Q=0) residuals

SPACF, the cut off at the first 12-1 lags occurs at Lag 9, and the cut off in multiples of 12 lags occurs at Lag 0, implying p=9 and P=0.

Based on the cutoff, the new set of possible sarima arguments are p=9, q=9, P=0, Q=0. Similarly we can derive another 2 competing models:

1. sarima(p=9, d=0, q=6, P=2, D=1, Q=2)
2. sarima(p=0, d=0, q=15, P=2, D=1, Q=2)

For both of the competing model listed above, sarima(p=9, d=0, q=6, P=2, D=1, Q=2) and sarima(p=0, d=0, q=15, P=2, D=1, Q=2), they

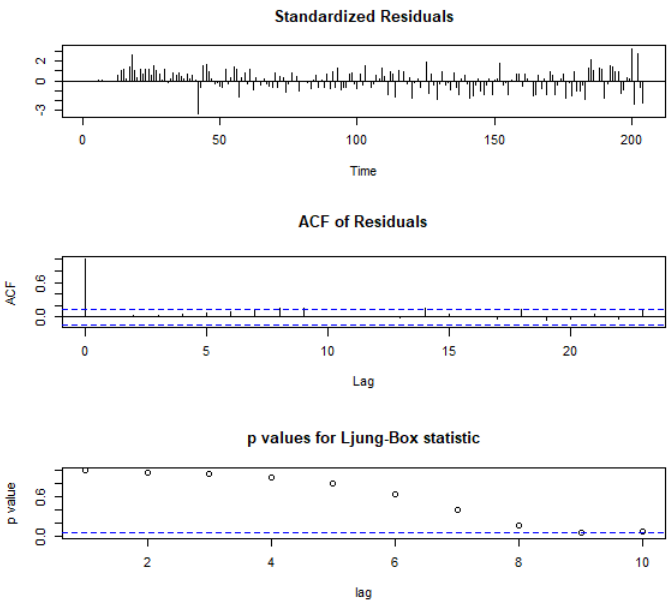

Figure 30: Fit2 Improvement 1 sarima(p=0, d=0, q=6, P=2, D=1, Q=2) tsdiag result

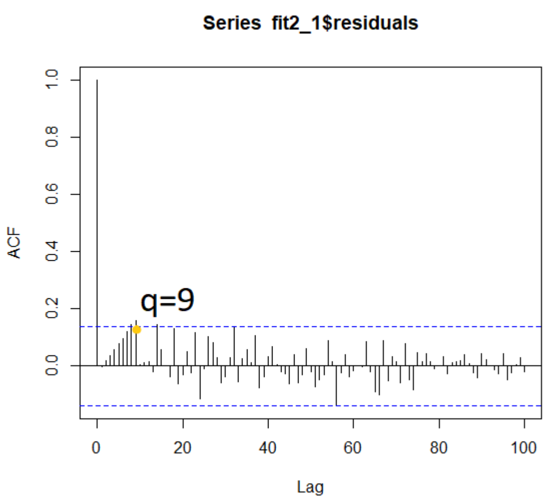

Figure 31: SACF Plot of Fit2 Improvement 1 sarima(p=0, d=0, q=6, P=2, D=1, Q=2) residuals

were already evaluated during Fit1's third improvement suggestion earlier in the report. As per our evaluation earlier, findings suggested that it passes adequate test with tsdiag outputs shown in Figure 25 and Figure 24 respectively. As described earlier, they yield an AIC of -520.96 and -484.51 respectively. We shall move on to the other possible improvements.

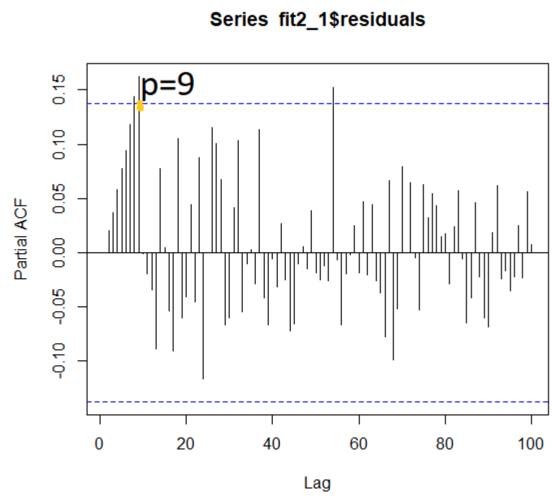

Figure 32: SPACF Plot of Fit2 Improvement 1 sarima(p=0, d=0, q=6, P=2, D=1, Q=2) residuals

5.2.2 Fit2's Improvement 2: sarima(p=4, d=0, q=5, P=4, D=1, Q=0). Applying the second improvement, model fitting sarima(p=4, d=0, q=5, P=4, D=1, Q=0), the results produced by the tsdiag test is as shown in Figure 33. The fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. This model will return us an AIC score of -518.46.

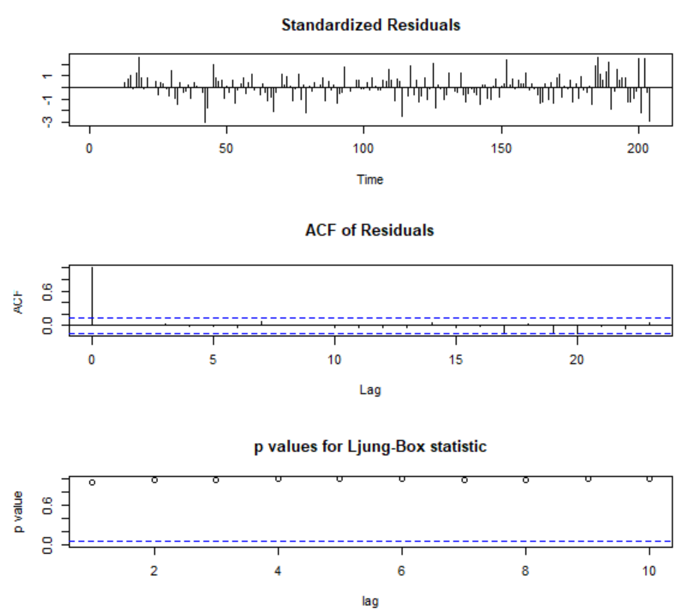

Figure 33: Fit2 Improvement 2 sarima(p=4, d=0, q=5, P=4, D=1, Q=0) tsdiag result

5.2.3 *Fit2's Improvement 3: sarima(p=0, d=0, q=6, P=4, D=1, Q=0).*

Applying the third improvement, model fitting sarima(p=0, d=0, q=6, P=4, D=1, Q=0) passes our adequate check as the results produced by the tsdiag test is as shown in Figure 34. The fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. This model will return us an AIC score of -425.38

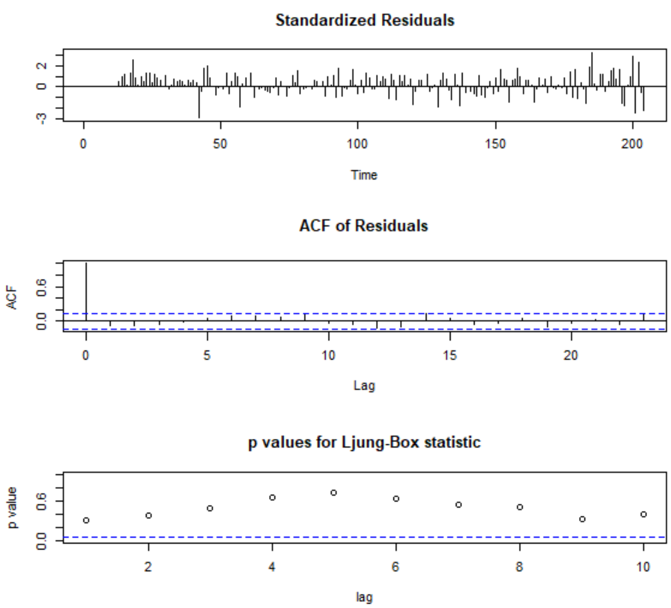

Figure 34: Fit2 Improvement 3 sarima(p=0, d=0, q=6, P=4, D=1, Q=0) tsdiag result

5.2.4 *Fit2's Improvement 4: sarima(p=4, d=0, q=5, P=2, D=1, Q=2).*

For the competing model listed above, sarima(p=4, d=0, q=5, P=2, D=1, Q=2) was already evaluated during Fit1's Improvement 2 suggestion earlier in the report. As per our evaluation earlier, findings suggested that it passes adequate test with tsdiag outputs shown in Figure 20. As described earlier, the SARIMA fitting yield an AIC of -519.37. We shall move on to the other possible improvements.

5.2.5 *Fit2 SARIMA((p=0, d=0, q=5, P=2, D=1, Q=0) Conclusion.*

Even though the original Fit2, SARIMA((p=0, d=0, q=5, P=2, D=1, Q=0) is not a valid fitting, we managed to come up with 2 additional valid improved fittings which are adequate and pass the tsdiag diagnostic check. The fittings as followed:

1. sarima(p=4, d=0, q=5, P=4, D=1, Q=0) with AIC score -518.46
2. sarima(p=0, d=0, q=6, P=4, D=1, Q=0) with AIC score -425.38

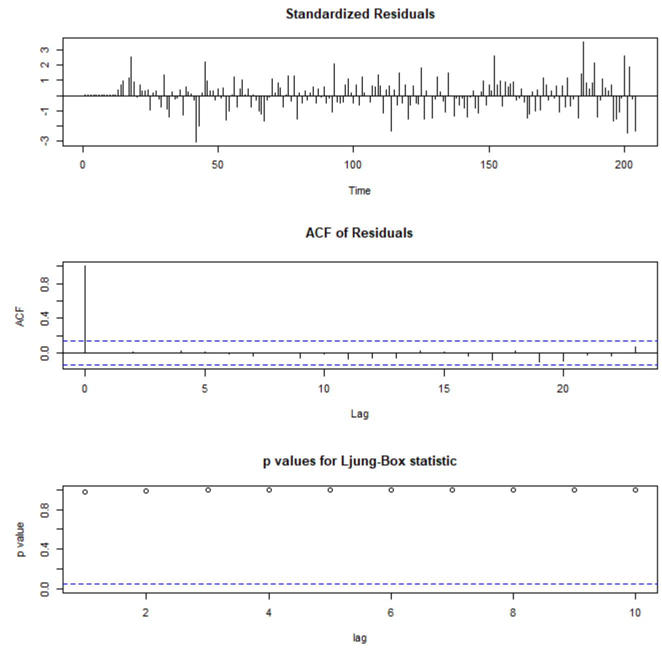

Figure 35: Fit3 sarima(p=10, d=0, q=0, P=2, D=1, Q=0) tsdiag result

5.3 Fit3: SARIMA((p=10, d=0, q=0, P=2, D=1, Q=0)

The tsdiag output shown in Figure 35, tells us that the fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. The model fitting of sarima(p=10, d=0, q=0, P=2, D=1, Q=0) return us an AIC score of -513.12.

5.4 Fit4: SARIMA((p=10, d=0, q=0, P=0, D=1, Q=2)

The tsdiag output shown in Figure 36, tells us that the fitting passes the adequate check as per the Ljung-Box Statistic Test as p-value is greater than 0.05 at all lag. The residuals are also randomized enough and we can also note that the residual ACF cut off instantly, implying that the residuals are White Noise. The model fitting of sarima(p=10, d=0, q=0, P=0, D=1, Q=2) return us an AIC score of -517.68.

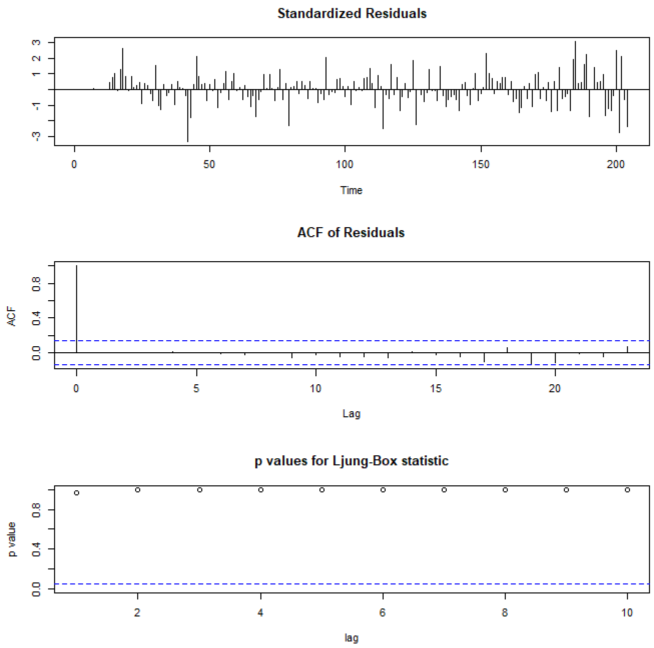

Figure 36: Fit4 sarima(p=10, d=0, q=0, P=0, D=1, Q=2) tsdiag

6 SARIMA MODELS SUMMARY TABLE

SARIMA models (p, d, q, P, D, Q)	AIC Score
(0, 0, 17, 4, 1, 4)	-499.0
(9, 0, 8, 4, 1, 4)	-511.74
(0, 0, 17, 0, 1, 4)	-490.55
(9, 0, 8, 0, 1, 4)	-518.11
(9, 0, 6, 0, 1, 4)	**-521.38**
(4, 0, 5, 2, 1, 2)	-519.37
(0, 0, 15, 2, 1, 2)	-484.51
(9, 0, 6, 2, 1, 2)	-520.96
(4, 0, 5, 0, 1, 4)	-520
(4, 0, 5, 4, 1, 0)	-518.46
(0, 0, 6, 4, 1, 0)	-425.38
(10, 0, 0, 2, 1, 0)	-513.12
(10, 0, 0, 0, 1, 2)	-517.68

Table 1: AIC Scores

Models that fail adequate check were discarded. Table 1 shows the AIC scores of all the models that pass adequate check. The lower the AIC score, the better. As such, the best AIC score of -521.38 is achieved by the sarima(p=9, d=0, q=6, P=0, D=1, Q=4) model. This model is the one selected for the forecasting.

7 SARIMA MODEL FORECAST

With the best SARIMA model selected, forecasting is performed by predicting 3 years ahead based on the model. Since the data that was input into the model was transformed using Box-Cox transformation, the prediction had to be inversely transformed to match the values of the original data. This was done using the InvBoxCox function, and is plotted in Figure 37. The upper and lower bounds of the forecast is also plotted.

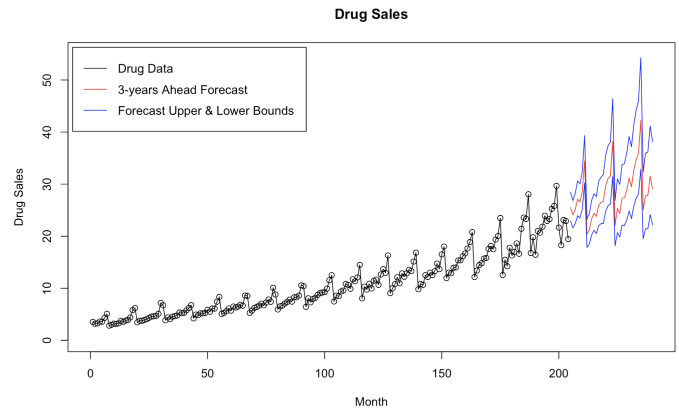

Figure 37: Drug Sales 3 years ahead Forecast as per SARIMA(9,0,6,0,1,4)

8 ADDITIONAL MODEL: HOLT-WINTERS

The Holt-Winters method is a commonly used technique for time series forecasting, particularly when dealing with data that exhibits seasonal patterns, such as sales data that spikes during the holiday season or weather data that follows a predictable pattern throughout the year.

For the Holt-Winters model, the time series data are first split into a training set and a test set. The training set contains the first 80% of the data, while the test set contains the remaining 20%. These sets are stored in y2_train and y2_test variables, respectively. The Add Model achieved RMSE of 1.8579 on Test set. The Multiply Model achieved RMSE 3.2586 on Test set. Both of which are acceptable fittings.

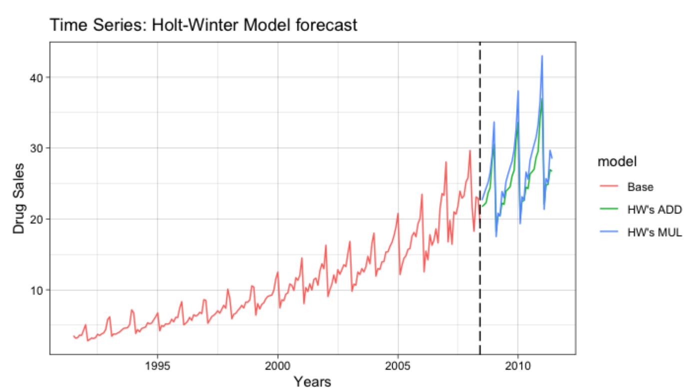

Figure 38: Holt-winter model forecast

9 CONCLUSION

The plot in Figure 39 shows the forecast values of the drug sales based on 3 different models, the best SARIMA model and together with the Holt-Winters Add and Multiply models.

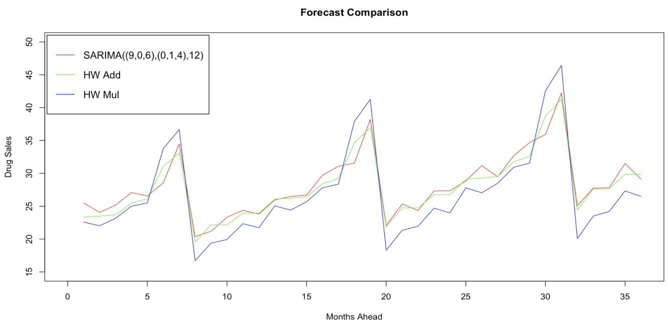

Figure 39: Forecast Comparison of the 3 types of model

In this further analysis, we compare the Holt-Winters and SARIMA models on a time series dataset. Both models were able to capture the seasonal and trend components of the data and provided reasonable forecasts.

Here are some difference we can note based on our experiment experience:

1. HW Model fitting is way faster. Some of the SARIMA fitting takes a long time when we run the fitting on our MacBooks used to conduct the experiment. HW Model fitting is usually instant. Most of the SARIMA with greater number (p, q, P, Q) arguments takes about 1 to 5 minutes.
2. HW Models can be broken into its trend, seasonal, and level components, this allows us to interpret how the model derives its forecasts and predicts, whereas for SARIMA, often we ended up with a list of coefficients for each of the AR and MA term, without understand why they are of such values, even with expert domain knowledge.
3. Our sarima is also tougher to fit as they were fitted through a tedious heuristic search. HW models were fitted automatically by functions provided by R libraries. We did not use auto.arima as it appear that the automation function does not work well to detect the seasonal component of this dataset, similar to the situation shown in Lecture. This result in the auto.arima models having very low AIC score relative to the one we fit heuristically. Therefore we omit using auto.arima.