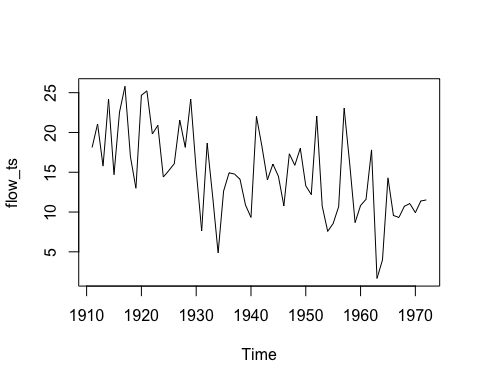
Quynh Tran

Stat 447 Final Project

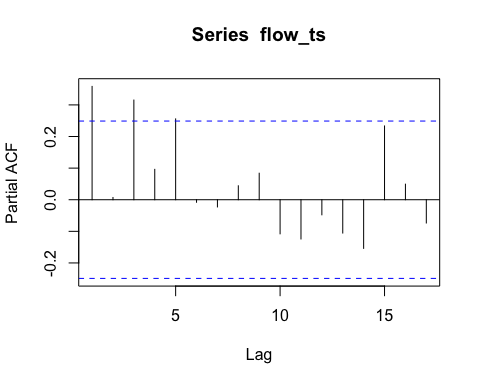
05/03/2023

# Annual flows for Colorado River 1911-1972.csv

The data “annual flows for Colorado River 1911-1972.csv” contains the annual for of Colorado river from 1911 to 1972. The dataset has two features: Year and Flow.

Chart

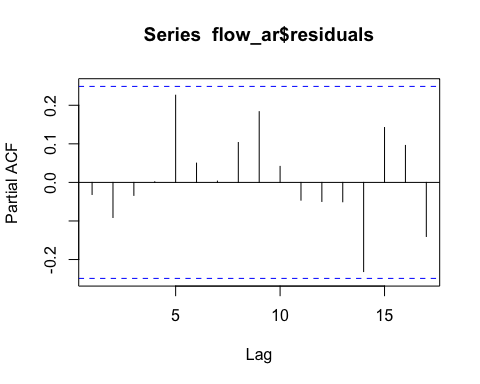
Description automatically generated



From the autocorrelation plot, the data looks somewhat not stationary enough, so we can conduct a Dickey-Fuller test to see if the time series is stationary or not. The null hypothesis is that a unit root is present. If a unit root is present, then p > 0, and the process is not stationary. Otherwise, p = 0, the null hypothesis is rejected, and the process is considered to be stationary.

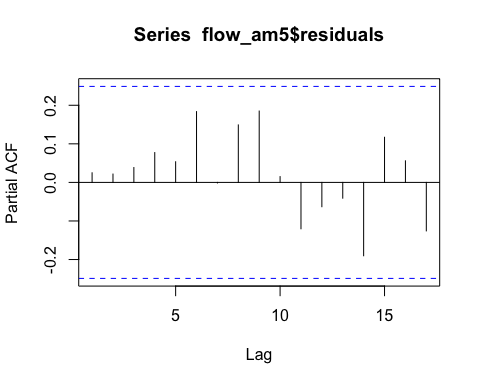
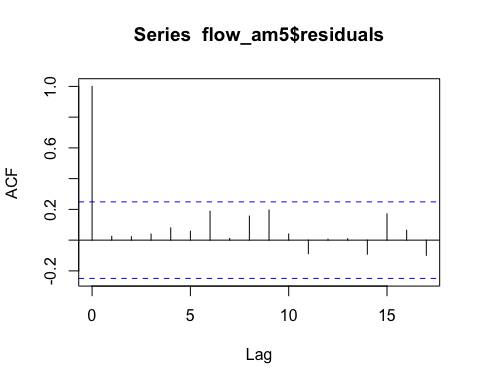
The Dickey-Fuller test resulted a p-value of 0.1008, which is very close to 0. This indicates that the time series is stationary.

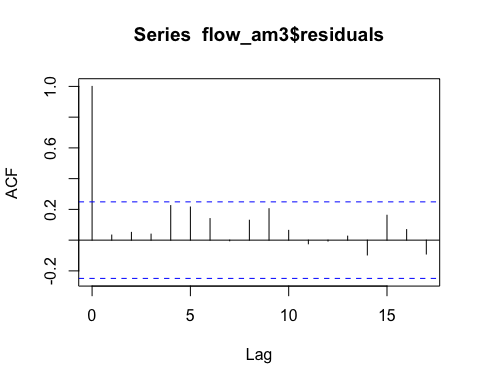
Chart

Description automatically generatedAdditionally, the ACF spikes die off and there are 3 significant spikes from PACF plot suggest an AR(3). The spikes from the ACF plot die off after lag 5, so I would also try to fit MA(5) and MA(3).

The ACF and PACF of the residuals from the AR(3) model suggest that we’ve accounted for most of the variance.

We then look at the ACF and PACF plots of the residuals from model MA(5) and MA(3):



Chart, box and whisker chart

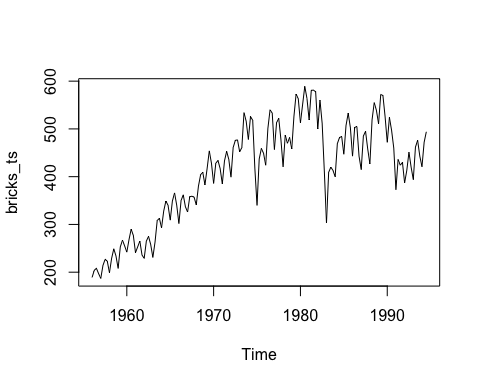
Description automatically generated

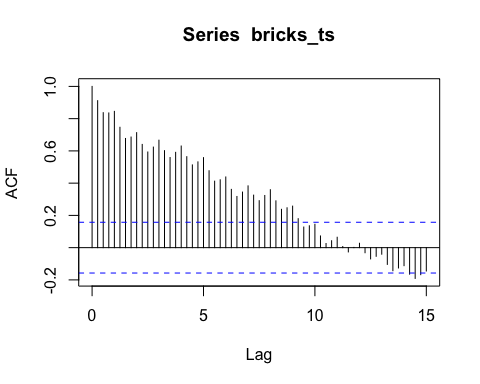
Similarly with models MA(3) and MA(5), there are no significant spikes in the residual plots.

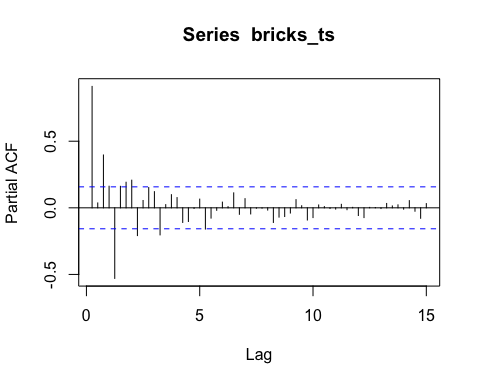
The three models also perform very similarly, with AIC score ranges around 380. The residual plots of the three models also show that we’ve accounted for most of the variance. Therefore, all AR(3), MA(3), MA(5) can be considered for this dataset.

# quarterly-production-of-clay-bricks.csv

The “quarterly-production-of-clay-bricks.csv” file contains the production of clay bricks every quarter from 1956 to 1994. The dataset contains 155 observations and has three 3 columns: year, qtr, and bricks



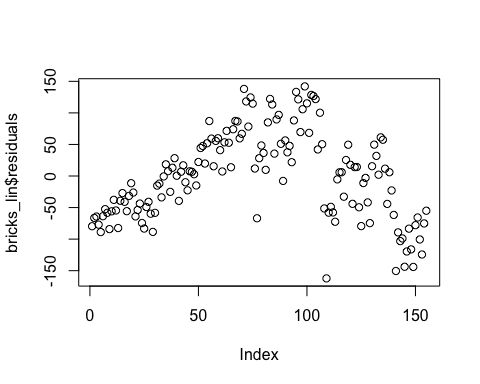


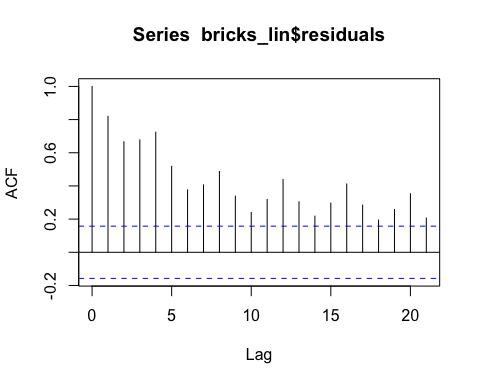


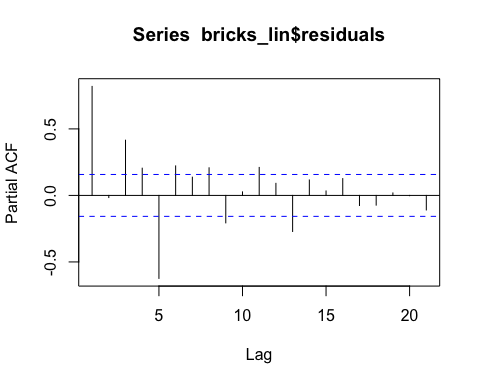
Notice that the data indicates an overall upward trend. It is reasonable that the production of clay bricks would increase over time and thus a regression model might be useful. However, if we focus on stochastic models, we will most likely need to utilize the first difference. We will look take the regression approach in this example. The trend could be deterministic (“time” is an explanatory variable) with the model possibilities of linear, quadratic or exponential or it could be stochastic where differencing will be appropriate.

The p-value of time trend (“tt”) are very small (0.0093) and with the visual evidence, it is reasonable to start using linear regression for this trend.

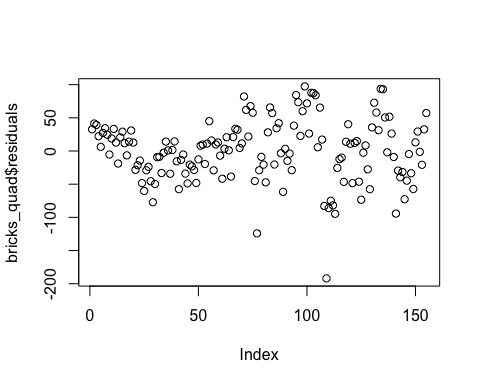
Let’s look at the plots of the residuals of the data set from the linear regression model:







The R-squared value from the linear regression is 0.58 that indicates 58% of the variation in the y values is accounted for by the x values. Addtionally, the ACF and PACF appear to show that we have explained only about half of the variation. Thus, we are moving to quadratic model.

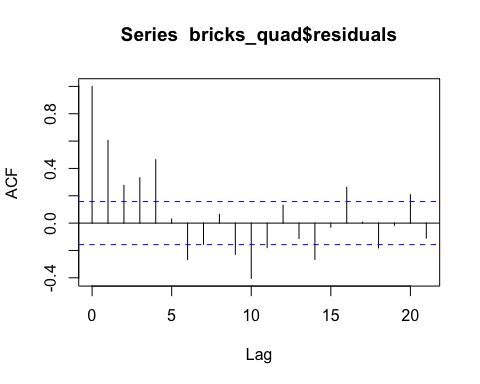


Below is a plot showing the residuals of the data after applying the quadratic model:

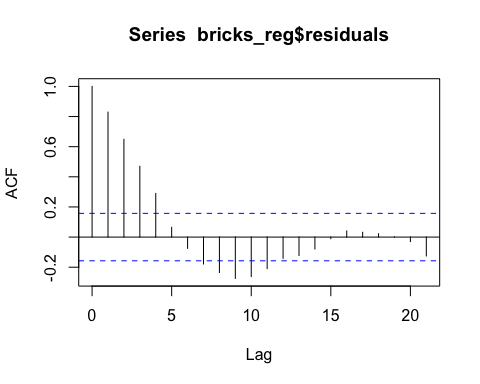
The R-squared of the quadratic model is 0.81, which means the model is able to explain up to 81% of the variance. The residual plot shows a very random pattern and the variability goes up as time increases.

The result is relatively high, but we will look at its ACF and PACF:

Chart, box and whisker chart

Description automatically generated

Notice that both of these show consistent spikes every 4 lags (quarters) indicating a seasonal effect. Since we have gone down the road of a regression model and it is reasonable that the bricks’ production would gave a seasonal effect, we will adapt our regression model to include the seasonality adjustment.

Chart

Description automatically generatedThe R-squared of the quadratic model with seasonality adjustment increases to 85.5%. The new model did not dramatically improve the r-squared value, however, when we look at the residuals, ACF and PACF of the residuals, we see that we have significantly improved the model:

We’ve gone about as far as we can with the regression model. Now we will model the residuals with a stochastic model. From the ACF and PACF we can see that we have removed the majority of the seasonal component and are left with AR(1) and MA(4) models (1 significant spike at lag 1 on the PACF and an ACF that dies off after lag 4. We will now fit AR(1), MA(4), and a combined ARMA(1,4) models to the residuals of the regression model.

Below are the plots obtained from AR(1):

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

MA(4) model:

Chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated

ARMA(1,4):

Chart, histogram

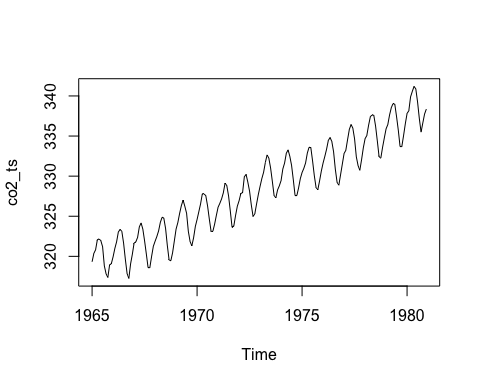
Description automatically generatedChart, box and whisker chart

Description automatically generated

Based on the ACF and PACF plots of the residuals from the three models, the mixed model - ARMA(1,4) performs the best as all the spikes are insignificant, which indicates no pattern left in the residuals.

# CO2 ppm Mauno Loan 1965-1980.csv

The csv file “co2-ppm-mauna-loa-1965-1980.csv” contains the amount of CO2 in Mauna Loa, Hawaii in ppm. The data was collected monthly from 1965 to 1980. The data set has 3 features: year, month, and co2.



The autocorrelation plot shows a seasonal trend and the curve is going upward over time. However, to verify that the data is stationary, I still conduct the Dickey-Fuller test to check its p-value.

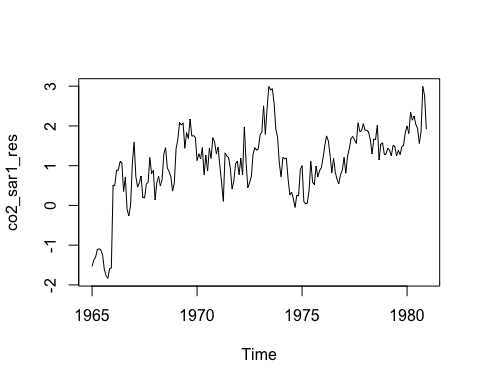
The p-value of resulted from the Dickey-Fuller test is 0.2438, which is relatively close to 0. Based on the autocorrelation plot and the small p-value, it is appropriate to conclude that the time series is stationary. Thus, we are moving on to analyzing the ACF and PACF plots:

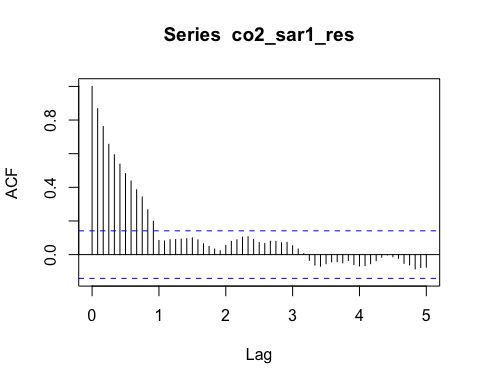
Chart, histogram

Description automatically generatedChart

Description automatically generated

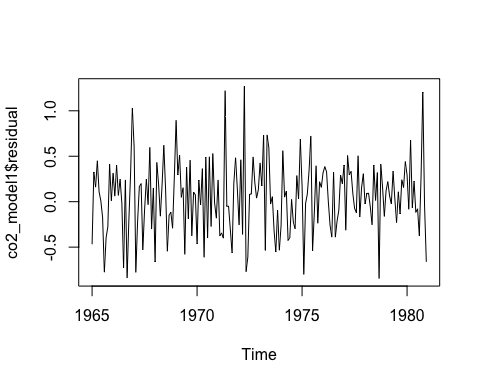
Tail-off is observed from the ACF plot that is an indication of AR model. From PACF, the two significant spikes are at lag 1 and lag 13 (12 lags apart), which indicates a potential SAR1 (seasonal moving average model of order 1).

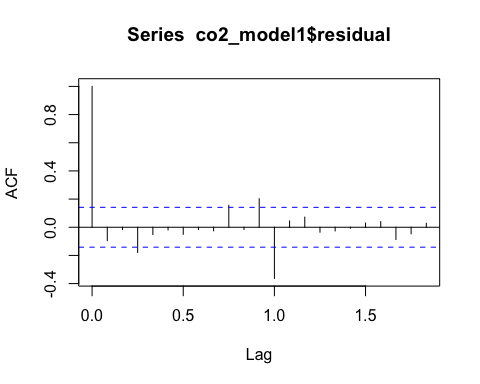


Chart

Description automatically generated

Here, there is an exponential decay from ACF plot that is also a sign an AR model may be appropriate. The PACF has two significant spikes at lag 1 and 13, indicating the residuals might follow an AR(1)model.



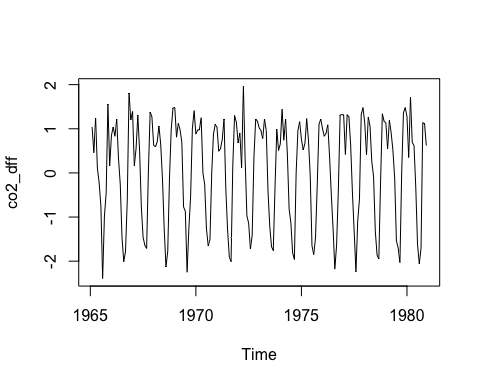
Chart, box and whisker chart

Description automatically generated

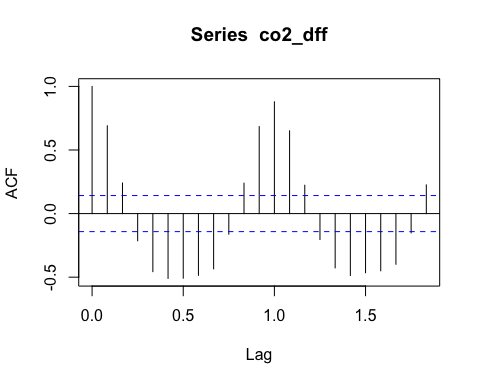
The ACF has one significant spike at lag 1. Similarly, PACF plot shows 1 significant spike at lag 1. This indicates that we account for some of the variation.

Now we will turn our attention to purely stochastic models to see if we can determine any other reasonable models. From our original plots, we know we need to account for trends. We do that by differencing the data to remove the trends.

Using differenced data, below is the autocorrelation plot:



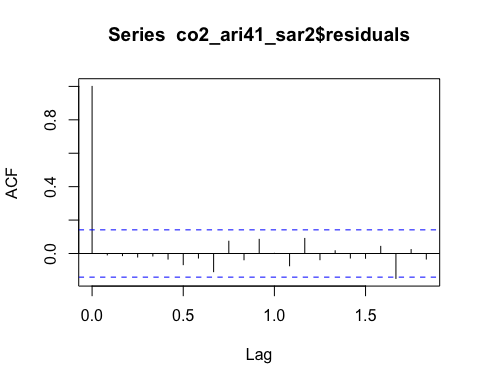
The differenced data looks very stationary.

Chart

Description automatically generated

We again see some concerning signs of conditional heteroskedasticity in the plot but will check again after we account for seasonality. The ACF and PACF show strong indications of seasonality (consistent spikes every 4 lags). Looking at the PACF, it looks like we have a significant spike at lag 4 indicating an AR(4) for the non-seasonal component and we have a marginally significant spike at lag 8 (2 seasons) indicating a possible seasonal AR(2) model. Thus, I will be investigating AR(2) and AR(4) for non-seasonal component and AR(2) for seasonal.

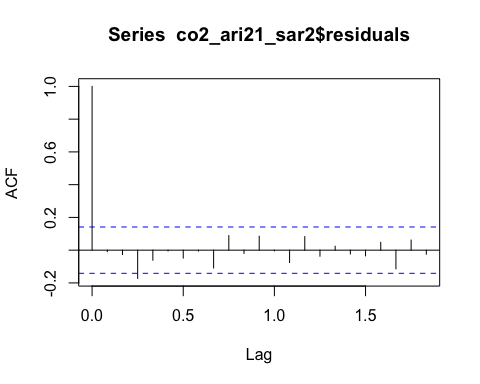
ACF and PACF of residuals of ARIMA(4,1,0) with seasonal AR(2):

Chart, box and whisker chart

Description automatically generated

The residuals are very much accounted for. However, there is one significant spike at the late stage from PACF.

Now, let’s look at ACF and PACF of residuals of ARIMA(2,1,0) with seasonal AR(2):

Chart, box and whisker chart

Description automatically generated

There is a significant spike at the very early stage of both ACF and PACF plots.

It looks like we have explained as much variation as possible with both model AR(4) with seasonality AR(2) and AR(2) with seasonality AR(2). Yet, since only significant spikes in ACF and PACF are at late lags and are marginally significant, model ARIMA(4,1,0) with AR(2) for seasonality is recommended over ARIMA(2,1,0) and SAR(1).

# annual-swedish-fertility-rates-1.csv

The “annual-swedish-fertility-rates-1.csv” file contains the annual fertility rate in Sweeden from 1750 to 1849.

Chart

Description automatically generated

Chart

Description automatically generatedChart, histogram

Description automatically generated

It is important to ensure that the data is stationary to justify a model before proceeding, so again, Dickey-Fuller test will be performed to double check this. However, based on the visualization, the data does not appear to be stationary and might require differencing.

After conducting the test, the obtained p-value is 0.82, which is large. Thus, the data is then differenced. Below are the autocorrelations plots of the differenced data:

Chart

Description automatically generated

The plot now appears to be more stationary.

Chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

The ACF shows that the spikes die off after lag 2 and there are 2 significant spikes in the PACF plot, so an MA(2) and AR(2) will be investigated. I will also study the mixed model: ARIMA(1,1) to see if it performs better.

MA(2) on differenced data’s residuals:

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

There are no significant spikes in the residual’s plots, which is a very clear evidence that no pattern can be identify and the model performs very well.

We now look at the visualizations from AR(2) model to see how it performs:

Chart

Description automatically generatedChart

Description automatically generated

The model seems to be less efficient for this data as some pattern can still be identified from the residuals. The ACF still shows significant spike at lag 3 and PACF shows significant spike at lag 3.

Now let’s look at the mixed model ARMA(1,1):

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Even though in most of the case, the mixed model tends to work better, ARMA(1,1) does not seem to be the best model. There are still some patterns from the residuals plots.

MA(2) model appears to be the winner among the three models that were investigated based on the visualizations.

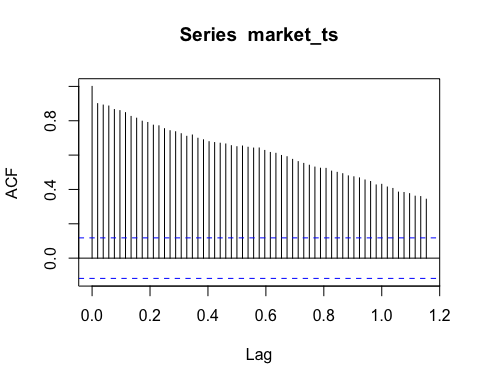
# weekly-market-share-data-of-cres 1958-1963-1.csv

The “weekly-market-share-data-of-cres 1958-1963-1.csv” file contains the weekly market share in Cres from 1958 to 1963.

Let’s glance at the autocorrelation, ACF, and PACF plots:

Chart

Description automatically generated



Chart, histogram

Description automatically generated

The autocorrelation plot indicates a soft upward trend in the data. This poses a concern that the data might appear to be stationary, but not enough to justify a model.

The Dickey-Fuller test results a p-value of 0.1732, which is relatively small. However, as mentioned, the shape of the autocorrelation plot calls for difference.

The data after being differenced:

Chart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated

The autocorrelation plot of the differenced data now appears to be more stationary. The ACF plot indicates the spikes die off after lag 1 and PACF shows 2 significant spikes, thus, we want to consider MA(1) and AR(2). However, the spikeat lag 4 is also slightly significant, so we will also consider AR(3).

MA(1) on differenced data:

Chart

Description automatically generatedChart

Description automatically generated

The residuals plots show no significant spikes, which means most of the variance has been accounted for.

Model AR(2) or ARIMA(2,1,0):

Chart

Description automatically generatedChart

Description automatically generated

There are still one or two significant spikes in both of the plots, so we will look at the next model, AR(3) with differenced data:

Chart, histogram

Description automatically generatedChart

Description automatically generated

This model also performs pretty well as no pattern is left in the residuals.

Based on the residuals plots, MA(1) and AR(3) are tie and we would not want to use AR(2) as it could not account for most of the variance. To choose one between MA(1) and AR(3), we can look at their AIC scores. MA(1)’s AIC is -795 and AR(3) is -906. We can see that AR(3) model yields a smaller AIC, thus, can be recommended as the optimal model.