香港大學

THE UNIVERSITY OF HONG KONG

CO2 EMISSIONS IN THE U.S.

Time Series Analysis

STAT4601 Final group project
2022-2023 semester 1

Name	UID
Cheung Ho Ting	3035688289
Emad Akhras	3035662962
Ng Chiu Fai	3035686956
Cheung Hiu Tung	3035694525

Abstract

Over the past few decades, climate change has become a worldwide phenomenon. The significant increase in the levels of greenhouse gases in the atmosphere, including but not limited to CO_2, have led to a rise in the temperature and eventually global warming. In this report, we model CO_2 emissions in the U.S. using data prepared by the U.S. Energy Information Administration. We examine the stationarity of the dataset and argue whether it requires further transformations. With reference to plots of ACF and PACF, we propose a few Seasonal ARIMA models to model-fit the data while justifying our choices along the way. Consequently, we perform model diagnostics on the models chosen previously by testing the normality of residuals in conjunction with Ljung-Box test. In addition, we explore expanding our models in an attempt to see if overfitting leads to better results, if any. Subsequently, we perform model selection and calculate parameter estimates. Ultimately, we aim to forecast the CO_2 volumes for last 12 months in the dataset and compare the observed against the predicted values. Lastly, we conclude the report with a summary of our findings and possible limitations and remedies.

Keyboards: CO_2 Emissions, Seasonal ARIMA

Table of Contents

1. **Introduction**

There is no question that environmental issues have become more prevalent in recent decades. The evidence is all around us – from melting glaciers and vanishing polar ice caps, to more extreme weather patterns and devastating natural disasters. And while there is still much debate over the root cause of these problems, there is a growing consensus that human activity – particularly the burning of fossil fuels – is playing a major role. An increase in the levels of CO_2 in the atmosphere is an inevitable consequence as such. Therefore, reducing CO_2 emissions must be a key priority for governments and organisations around the world in order to alleviate extreme weather conditions.

As global citizens of the world, we realize that a great deal of effort has been put towards solving the abovementioned issue in recent years. To mark our contribution, our group aims through this project to study, analyze, and predict the trend in the volume of CO_2 emissions in the U.S. The source of our data is the U.S. Energy Information Administration[1] where we extracted monthly data of CO_2 volumes measured in Million Metric Tons from January 1973 until July 2022. The dataset is comprised of CO_2 emissions from multiple industrial sectors such as Coal Electric Power, Natural Gas Electric Power, Distillate Fuel including Kerosene-type jet fuel, Oil Electric Power, Petroleum Coke Electric Power, Petroleum Electric, Residual Fuel Oil Electric Power, Geothermal Electric Power, and Non-Biomass Waste Electric Power.

Having briefly introduced the background of the dataset, the remainder of this report is arranged as follows. In section two, we analyze the stationary element of our time-series by drawing inferences from the time as well as transformed plots. In section three, we fit seasonal ARIMA models to the data and provide specifications with regards to the parameters of the seasonal as well as the ARIMA part. This is followed by section four in which we perform model diagnostics to assess the adequacy of the chosen models and estimate their parameters. In section five, a one-year forecast is conducted on the last twelve entries in the time series; that is from August 2022 until July 2023. Lastly, section six contains concluding remarks.

[1] (Administration, December 2021)

2. **Stationarity**

It is no surprise that the emergence of COVID-19 had dire ramifications on the world economy, and energy-related CO_2 emissions. Prior to that, however, CO_2 emissions in the U.S have generally declined particularly over the last decade. From *figure 1*[2], we see that since peaking in 2007, the magnitude of decline in 2020 was biggest than all previous years in percentage and absolute terms. This is largely due to a dramatic decline in demand for energy as a consequence of a global economic shock.

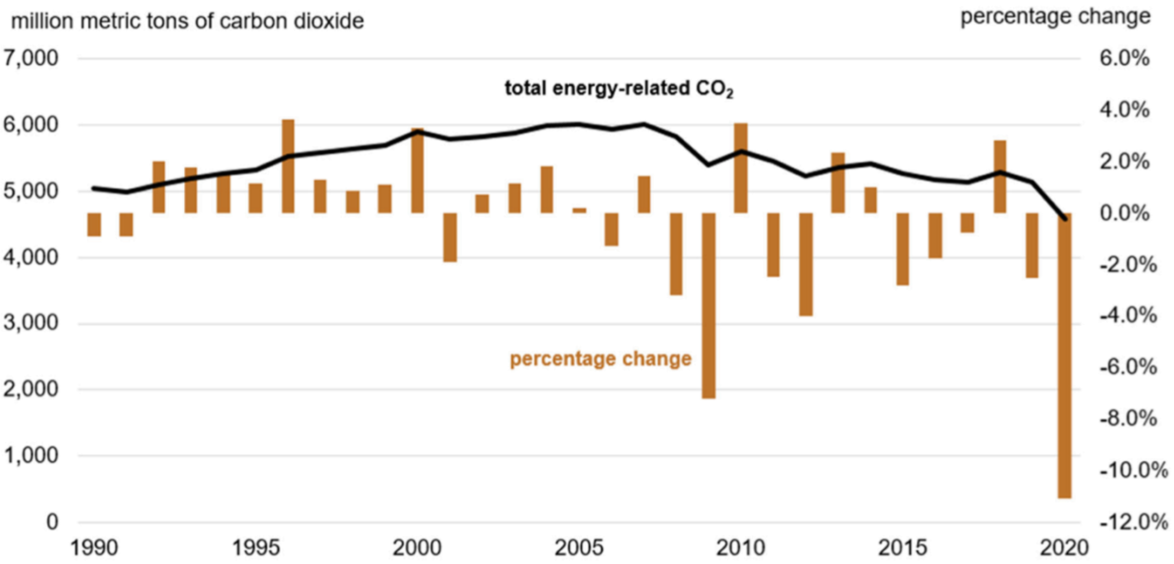


Figure 1: Annual emissions of energy-related carbon dioxide

The time plot of the data is presented in *figure 2*. Upon inspection, we notice an apparent increase in the variance towards the end of the studied period in comparison with the beginning. This pattern persisted in the log transformed plot of the data in *figure 3*. Another interesting observation is the upward then downward trend in both the original and log transformed plots. That, in conjunction with the slow-decaying pattern in the sample ACF plot in *figure 4*, provide clear indications that our time series is non-stationary. Besides, the sample ACF plot is characterized by a regular wave-like pattern which signals the existence of significant seasonal autocorrelations.

[2] (Administration, December 2021)

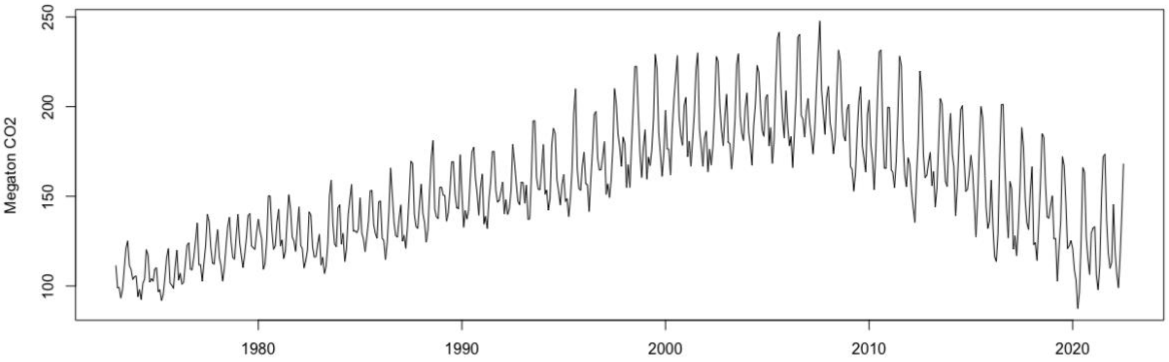


Figure 2: Time plot of the original data

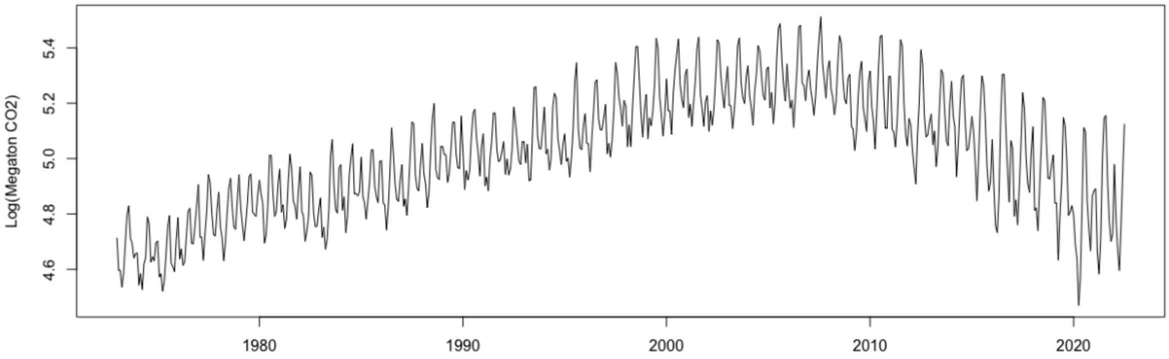


Figure 3: Time plot of log transformed data

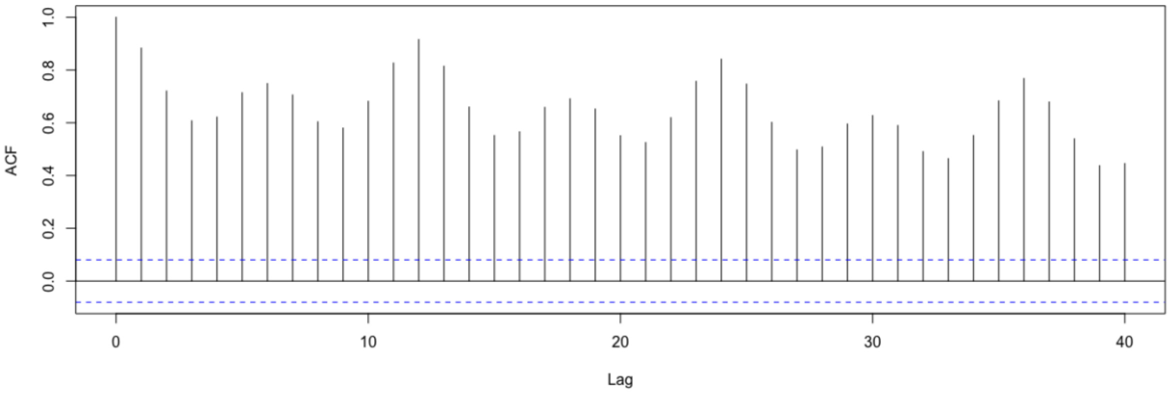


Figure 4: Sample ACF for log transformed data

Having concluded that our time series is non-stationary in mean, the next step is to inspect it further after taking a first-order difference. As a result, the time plot in *figure 5* doesn't exhibit any upward or downward trends. All types of ADF test at all lags produce p-values that are less than 5% significant levels which provides evidence to reject the null hypothesis that our time series is non-stationary. In addition, although the sample ACF plot in *figure 6* supports our claim that the trend has been smoothened, it reveals a strong recuring seasonality effect at lags that are multiples of 12, i.e., 12, 24, 36, etc.

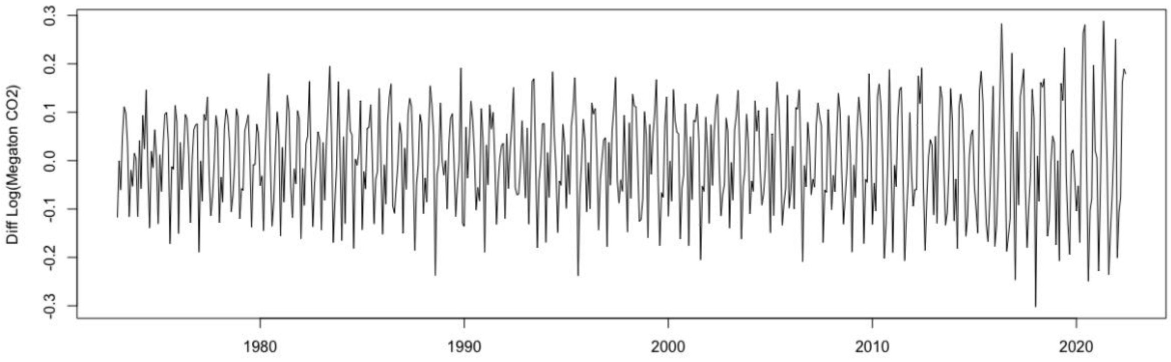


Figure 5: First Order Difference of log transformed data

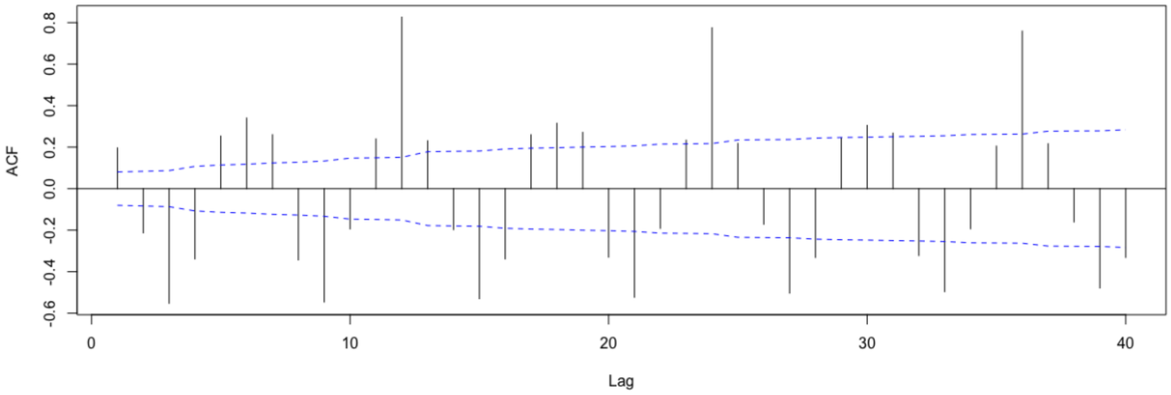


Figure 6: Sample ACF for first difference log transformed data

The above observation suggests that a seasonal difference with 12 periods may enable parsimonious modelling. *Figure 7* shows the sample ACF after one seasonal difference with 12 periods. The ACF is significant until lag 7 which will be discussed further in section 3. It is worth noting that the sample PACF are difficult to explain, hence we rely in our analysis of identifying suitable and tentative models on sample ACF. *Figure 8* illustrates the time plot of the log transformed data after taking a first-order difference as well as a seasonal difference with 12 periods. The ACF plot of that is shown in *figure 9.* Most of the seasonal correlations have been eliminated except the one at lag 12 which is still significant. Note that in most time plot figures we observe that the variance of carbon emissions in recent years is abnormally higher than that in previous years. This might have been due to reasons related to COVID-19 that are highlighted in the introduction. In retrospect, this may affect the stationarity of the data and hence the adequacy of the models despite the different transformations performed earlier. However, we should still proceed with SARIMA models because most of the data is stationary, and we believe that the variance of observations in the future should be less abnormal and more consistent.

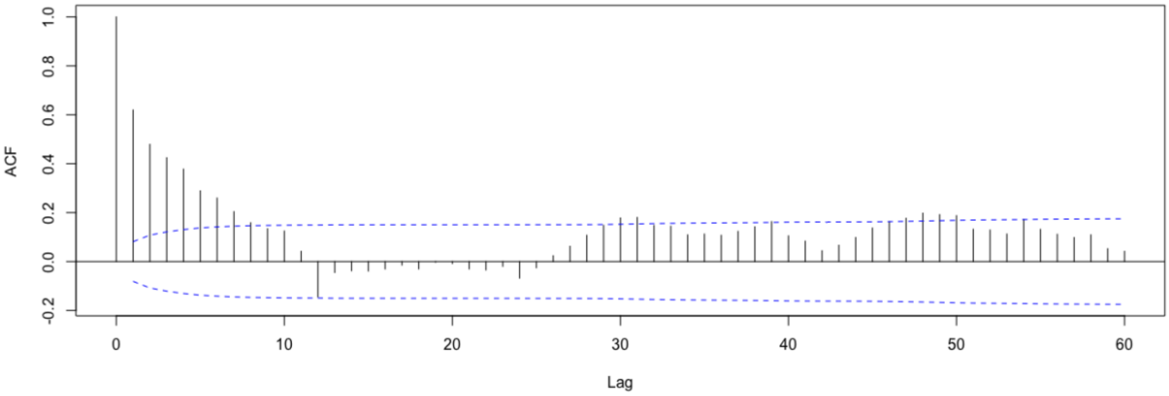


Figure 7: Sample ACF for one seasonal difference log transformed data

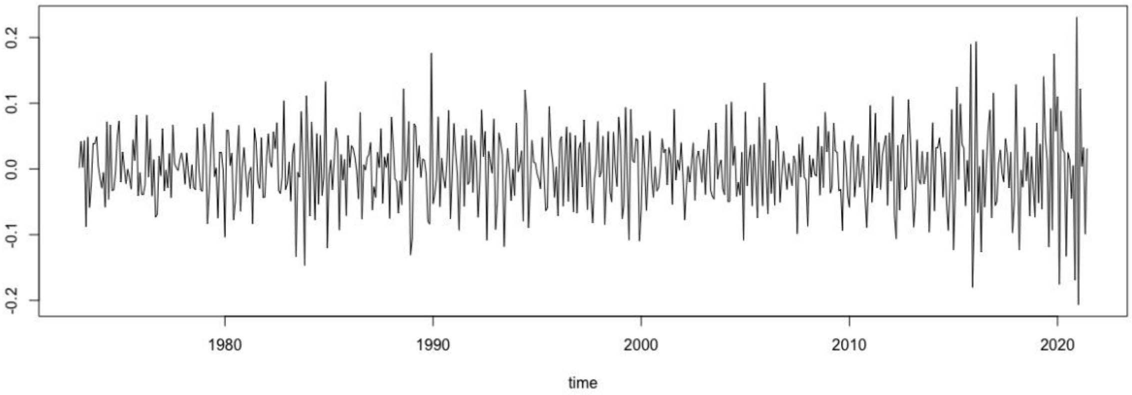


Figure 8: First order and seasonal diff of log transformed data

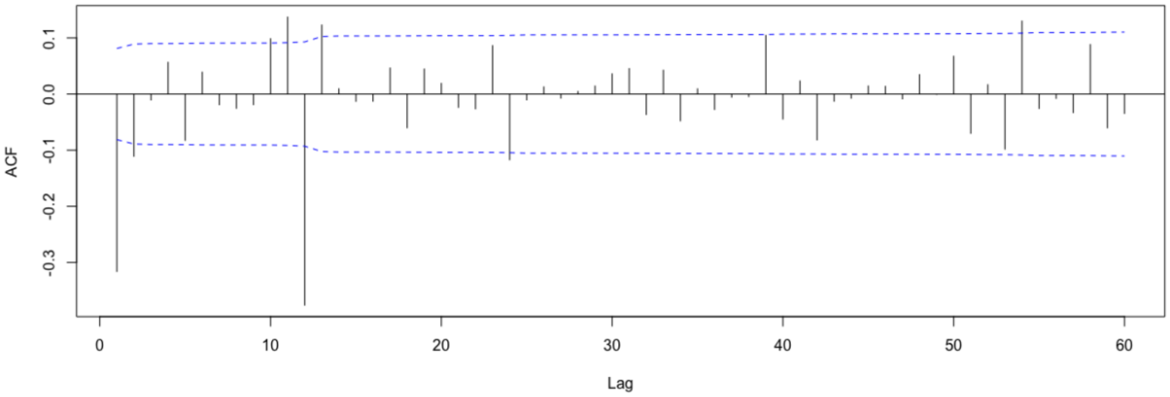


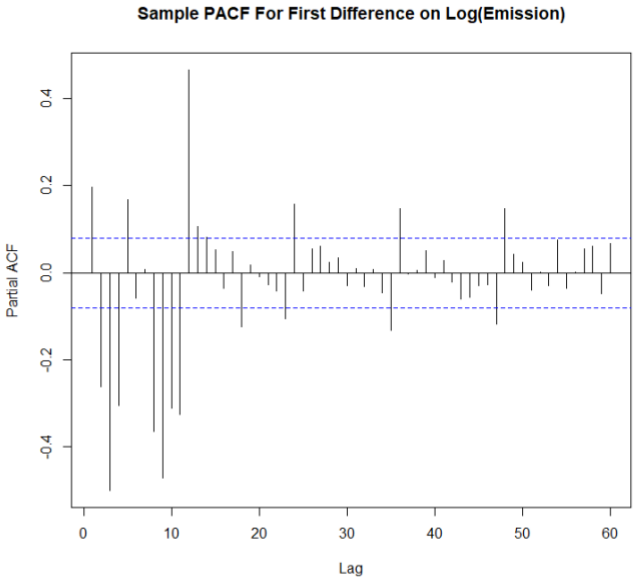
Figure 9: Sample ACF of first order and seasonal difference

3. **Model Specification**

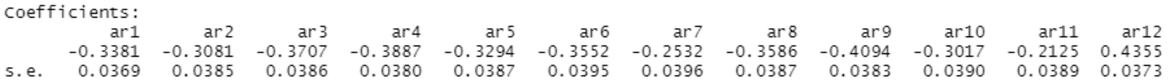
This section describes how we reach our selected model candidates by analyzing the patterns of autocorrelations in the data and residuals from smaller models. We have two model candidates in total, which would be discussed in detail in Section 3.1 and 3.2 respectively. Models are fitted using the *arima* function from the TSA package in R.

3.1. **Model 1**

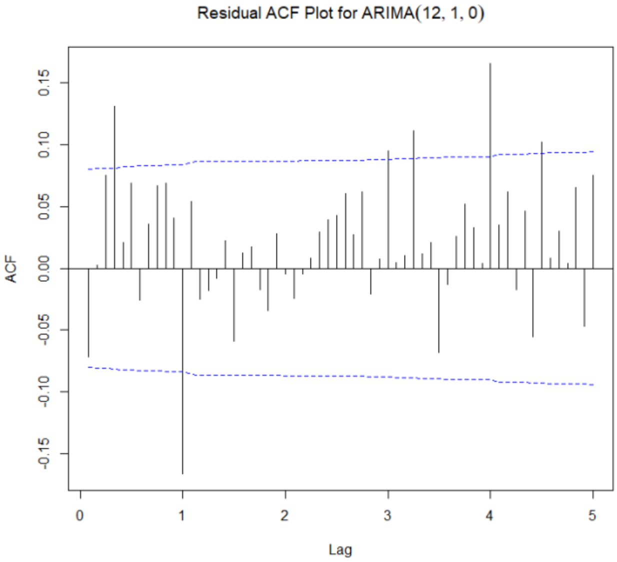
From the PACF plot of the log CO_2 emission after taking once differencing, we observe that the partial autocorrelations at the first 12 lags are mostly significant, with some little spikes in later lags possibly due to seasonality. Since the PACF at the first 12 lags are far more significant than the PACF which are caused by seasonality at later lags, we first account for the PACF from lag 1 to lag 12 and employ an $ARIMA(12,1,0)$ model as a base model.



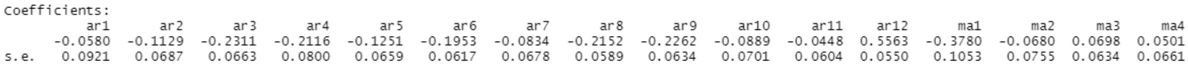
After fitting the model, we observe that all coefficients are significant with reasonable standard errors. There is a significant residual ACF at lag 4 and lag 12. We suspect that the significant ACF at lag 12 is due to seasonality. The significant residual ACF at lag 4 suggests that we may raise the MA parameters in our model to explain it. However, it seems difficult to observe any interaction from the seasonal residual ACF. Hence, we decide to neglect the seasonal ACF at lag 12 as well as later lags and proceed with an $ARIMA(12,1,4)$ model.



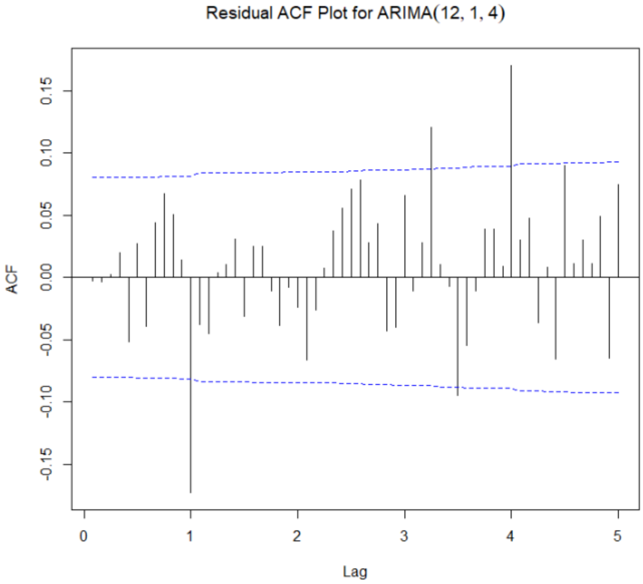
```
Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8      ar9     ar10     ar11    ar12
      -0.3381  -0.3081  -0.3707  -0.3887  -0.3294  -0.3552  -0.2532  -0.3586  -0.4094  -0.3017  -0.2125  0.4355
s.e.   0.0369   0.0385   0.0386   0.0380   0.0387   0.0395   0.0396   0.0387   0.0383   0.0390   0.0389  0.0373
```



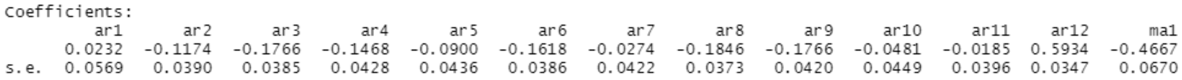
Having fitted the $ARIMA(12,1,4)$ model, there are now no significant residual ACF at lag 4 anymore, as desired. We notice that the coefficients for *ma2*, *ma3*, and *ma4* are all insignificant, so we decide to reduce the number of parameters to $ARIMA(12,1,1)$.



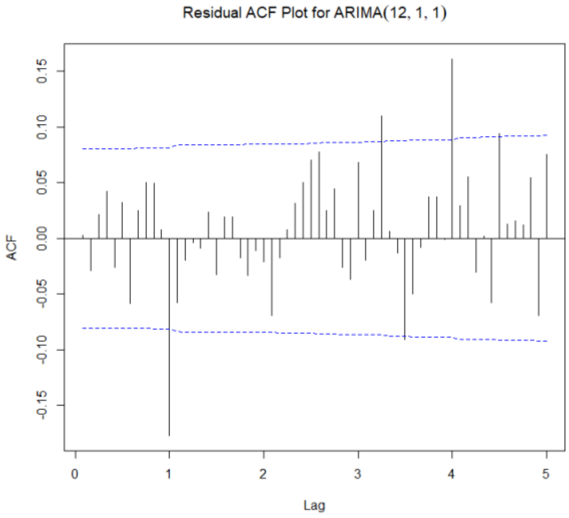
```
Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8      ar9     ar10     ar11    ar12      ma1      ma2     ma3     ma4
      -0.0580  -0.1129  -0.2311  -0.2116  -0.1251  -0.1953  -0.0834  -0.2152  -0.2262  -0.0889  -0.0448  0.5563  -0.3780  -0.0680  0.0698  0.0501
s.e.   0.0921   0.0687   0.0663   0.0800   0.0659   0.0617   0.0678   0.0589   0.0634   0.0701   0.0604  0.0550   0.1053   0.0755  0.0634  0.0661
```



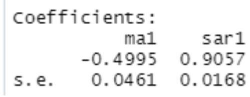
Although we have reduced the number of parameters, the model performance remains similar to previous model in terms of the residual ACF plot. We notice the coefficients from *ar1* to *ar11* are now less significant. It suggests that we should try model refining and employ a seasonal ARIMA model by setting all coefficients from *ar1* to *ar11* to zeros. We then proceed to observe the performance of such model, namely $SARIMA(0,1,1) \times (1,0,0)_{12}$.



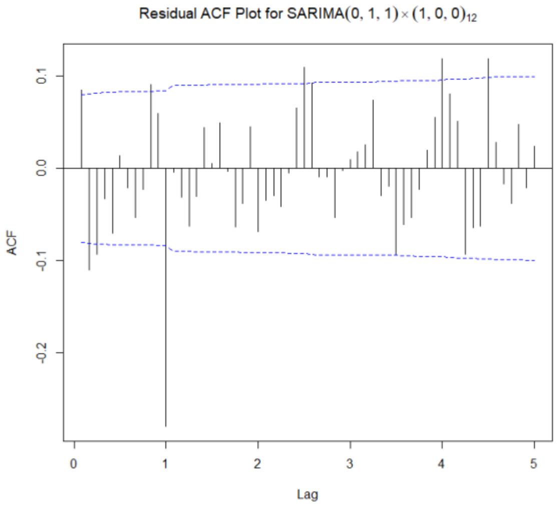
```
Coefficients:
         ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8      ar9     ar10     ar11    ar12      ma1
      0.0232  -0.1174  -0.1766  -0.1468  -0.0900  -0.1618  -0.0274  -0.1846  -0.1766  -0.0481  -0.0185  0.5934  -0.4667
s.e.  0.0569   0.0390   0.0385   0.0428   0.0436   0.0386   0.0422   0.0373   0.0420   0.0449   0.0396  0.0347   0.0670
```



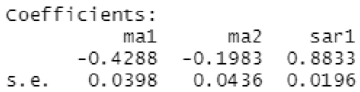
From the residual ACF plot, we observe slightly significant residuals autocorrelations before lag 12. This suggests that our model is still not adequate, as we have massively dropped some coefficients from the previous model. However, we are confident about our previous argument that the previous model is too big, so we decide to increase the parameters with small increment and observe the performance. This motivates us to try fitting an $SARIMA(0,1,2) \times (1,0,0)_{12}$ model.



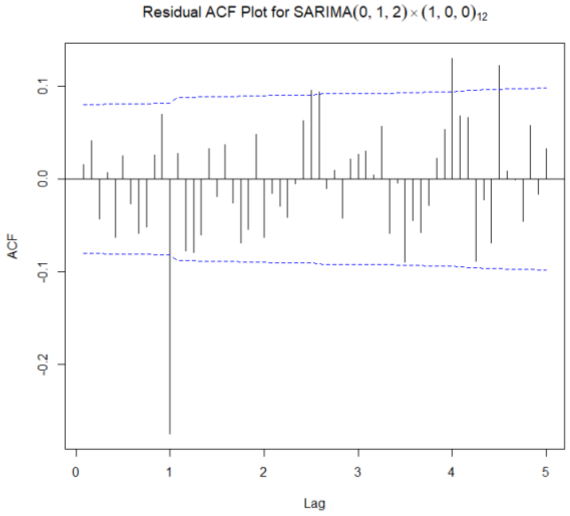
```
Coefficients:
          ma1    sar1
      -0.4995  0.9057
s.e.   0.0461  0.0168
```



We can see that the residual ACF are now insignificant in most lags excluding lag 12, which is caused by seasonality. All coefficients are quite significant with small standard errors. Unfortunately, this undesired significant ACF at lag 12 is very difficult to get rid of. We will address this issue in section 4 later together with model diagnostics. Despite the undesirable ACF at lag 12, we think this model is good enough and seems adequate, so this model is selected as our first candidate for model selection.

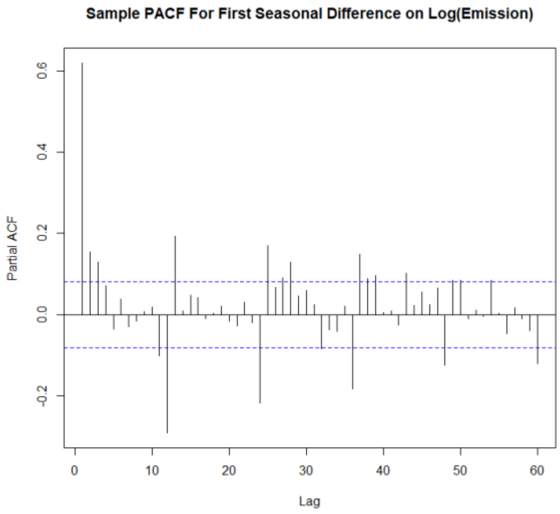


```
Coefficients:
          ma1      ma2    sar1
      -0.4288  -0.1983  0.8833
s.e.   0.0398   0.0436  0.0196
```

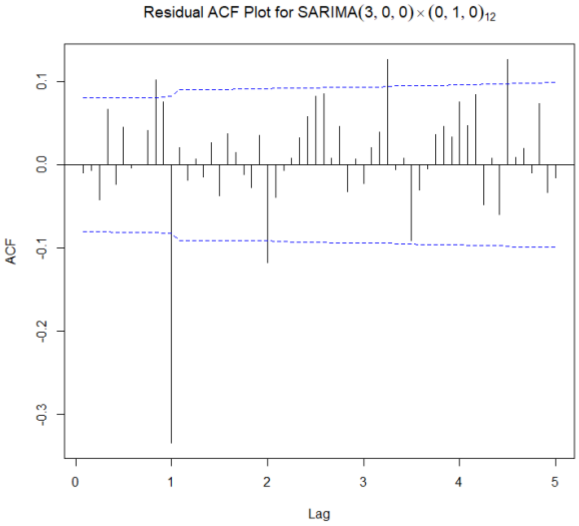


3.2. **Model 2**

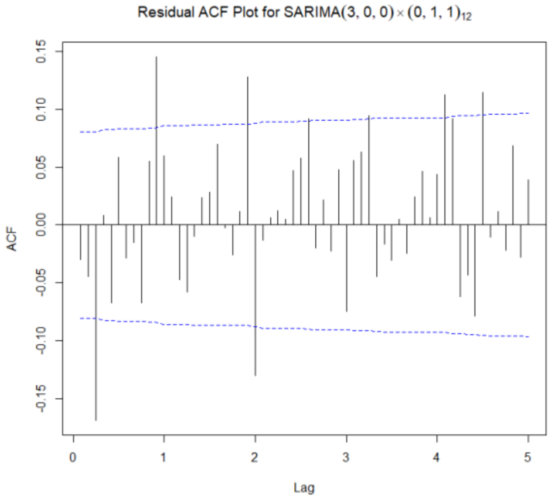
In the previous section, we proposed a model based on the data after taking log and once differencing. In this section, we consider transforming the data by taking log and once seasonal differencing with period 12. From the partial autocorrelation plot of the transformed data, we notice significant PACFs from lag 1 to lag 3, followed by periodic significant PACFs in later lags which decay relatively slowly. Since it is difficult to interpret seasonal effects from the PACF plots, we would like to neglect the seasonal effects and fit a model to explain autocorrelation in the common parts, that are present in the first 3 lags. Under this motivation, it is natural to consider an $SARIMA(3,0,0) \times (0,1,0)_{12}$ model.



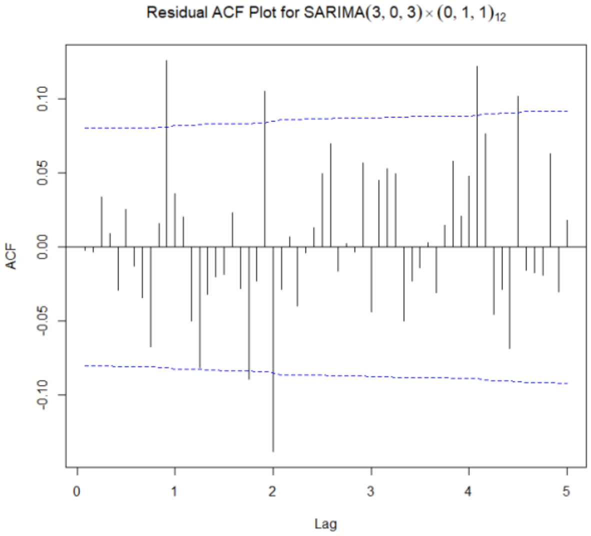
From the residual ACF plot obtained from fitting the abovementioned model, we observe that the autocorrelations in most lags are insignificant. The ACF at lag 12 is exceptionally significant with some marginally significant ACFs in later lags. The overwhelmingly significant ACF at lag 12 motivates us to use an $SMA(1)$ model to explain it. Combining this with our model, we have an $SARIMA(3,0,0) \times (0,1,1)_{12}$ model.



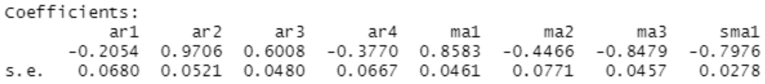
Although the residuals ACF plot shown below does not suggest an improvement, the AIC (Akaike Information Criterion) value of this model has decreased from -1823 to -2000 compared to the previous model. This suggests that this model may have a better fit in general. We notice that there is significant ACF at lag 3. The ACFs around the lag 12, 24, 36, and so on, may be slightly significant due to the interaction between some effects in the common part and the seasonal part. We decide to fit a bigger model and observe its performance, namely $SARIMA(3,0,3) \times (0,1,1)_{12}$



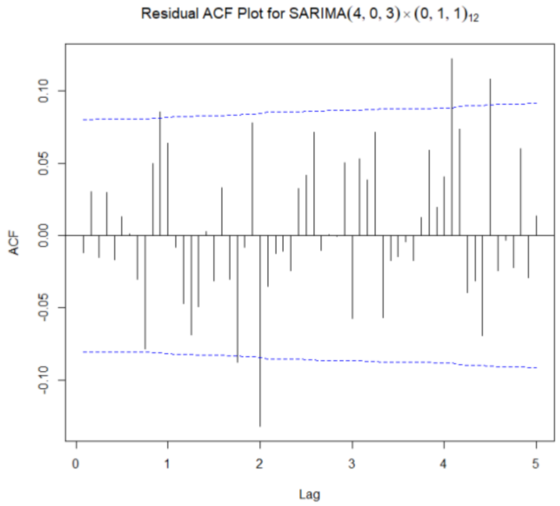
Having fitted the $SARIMA(3,0,3) \times (0,1,1)_{12}$ model, we plot its residuals ACF. We notice that the significant ACF at lag 3 is not present now, as desired. However, the significant ACFs at lags 12 and 24 are still present due to unexplained seasonality. It is a clear sign that our model may not be adequate, which motivates us to try incrementing the number of parameters.



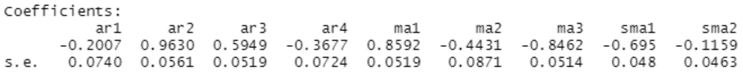
After trying to increment different parameters by one, we notice that an $SARIMA(4,0,3) \times (0,1,1)_{12}$ model can explain the seasonality effects fairly better as the residual ACF at lag 12 is not significant anymore. It is also important to note that all coefficients are considerably significant with very small standard errors. It seems that the slightly significant residual ACF after the fourth period may be caused by noise or outliers. Hence, it remains to explain the significant residual ACF at lag 24, which drives us to try an $SARIMA(4,0,3) \times (0,1,2)_{12}$ model.



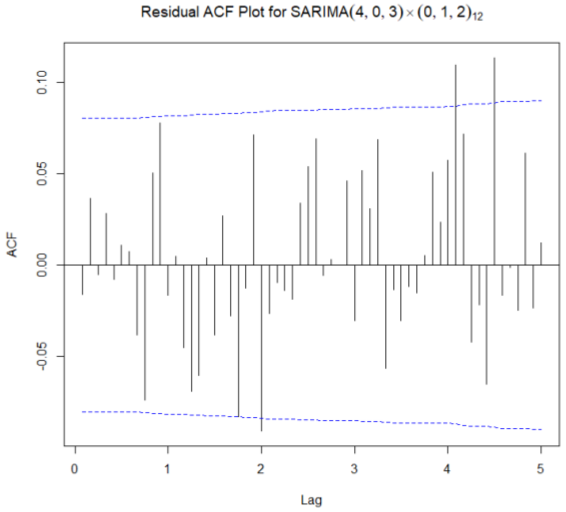
```
Coefficients:
          ar1     ar2     ar3      ar4     ma1      ma2      ma3     sma1
      -0.2054  0.9706  0.6008  -0.3770  0.8583  -0.4466  -0.8479  -0.7976
s.e.   0.0680  0.0521  0.0480   0.0667  0.0461   0.0771   0.0457   0.0278
```



We can see that the residual ACFs from the $SARIMA(4,0,3) \times (0,1,2)_{12}$ model are mostly insignificant, as desired. The coefficients are mostly significant with small standard errors. Hence, this model is selected as the final candidate. Details of the residual analysis for the selected model candidates, Model 1 and Model 2, will be discussed in the next section.



```
Coefficients:
          ar1     ar2     ar3      ar4     ma1      ma2      ma3    sma1     sma2
      -0.2007  0.9630  0.5949  -0.3677  0.8592  -0.4431  -0.8462  -0.695  -0.1159
s.e.   0.0740  0.0561  0.0519   0.0724  0.0519   0.0871   0.0514   0.048   0.0463
```



4. **Diagnostics**

As we have arrived at our models, we perform diagnostics on the models from two directions, residual analysis followed by overfitting. In the residual analysis part, we further analyze both the normality and autocorrelations of the residuals.

4.1. **Residual Analysis**

4.1.1. ***SARIMA*(0,1,2) x (1,0,0)$_{12}$**

Our first model is *SARIMA*(0,1,2) x (1,0,0)$_{12}$. We first check the normality of the residuals by its time plot, histogram, and the Q-Q plot. The residuals exhibit a constant mean and a relatively stable variance despite being slightly enlarged at the end of the timeline. Also, since the histogram has a good-looking bell shape, and most residuals lie on the indicator line of the Q-Q plot with only a few outliers at the ends, we propose that the residuals follow a normal distribution.

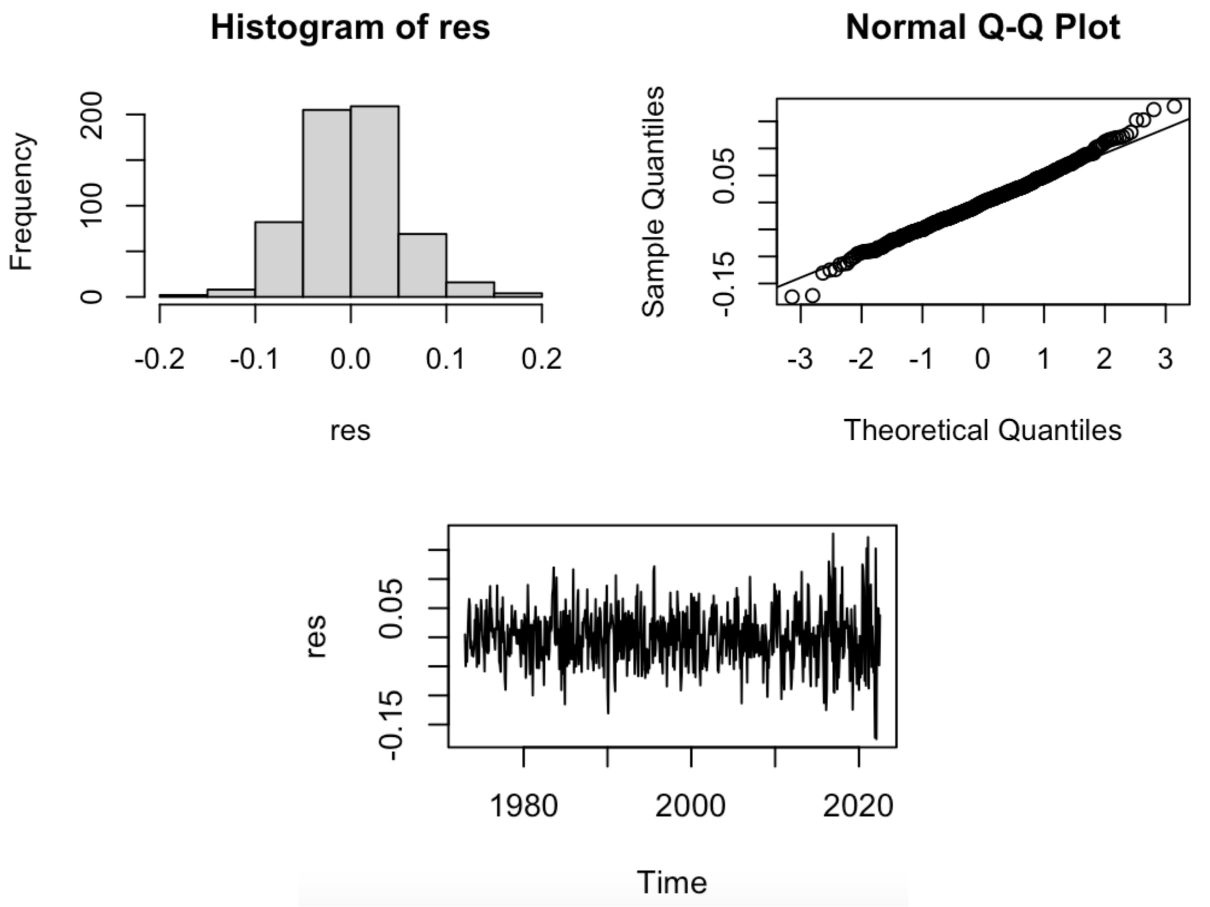
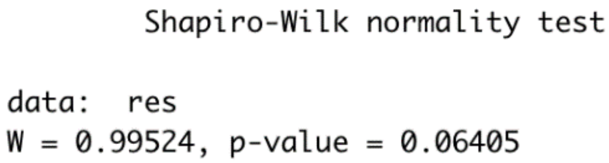


Figure 10: Histogram, Q-Q plot and time plot of residuals of SARIMA(0,1,2) x (1,0,0)$_{12}$

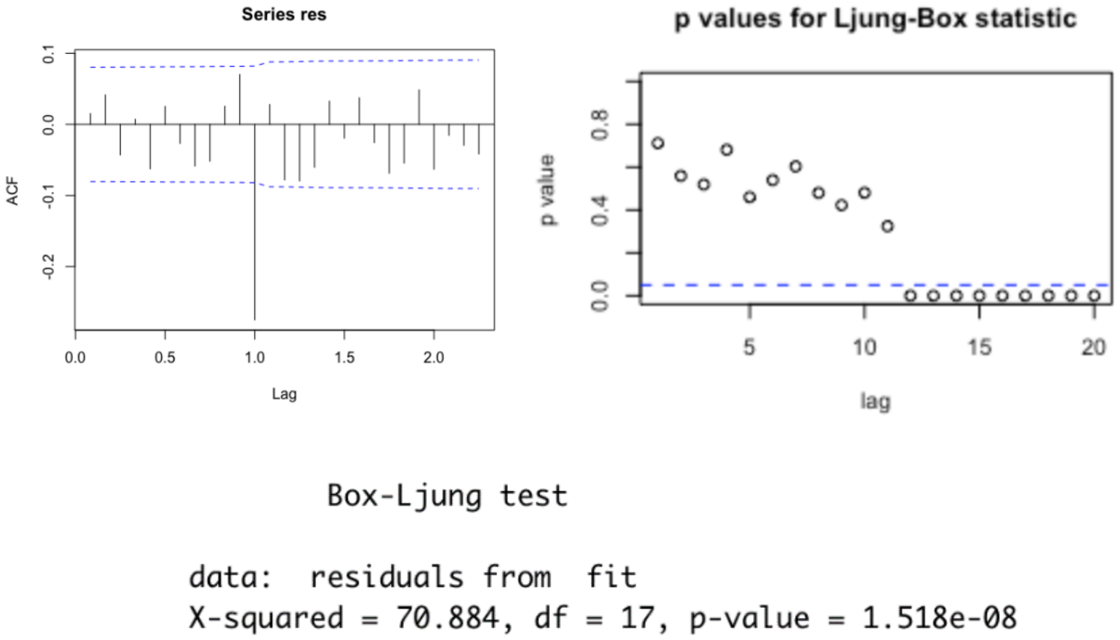
Then, we perform the Shapiro-Wilk normality test on the residuals. This standard normality test gives us a p-value of 0.06405, which is greater than 0.05, thus confirming the normality of the residuals.



```
            Shapiro-Wilk normality test

data:  res
W = 0.99524, p-value = 0.06405
```

Figure 11: Standard normality test for residuals of SARIMA(0,1,2) x (1,0,0)$_{12}$

We proceed to look for any possible autocorrelations of the residuals. Unfortunately, the ACF of the residuals is significant at lag 12, indicating that possibly some information is not included in the model. As a result, the Ljung-Box test also gives a p-value close to 0 starting from lag 12, and our models fails the autocorrelation test at high lags.



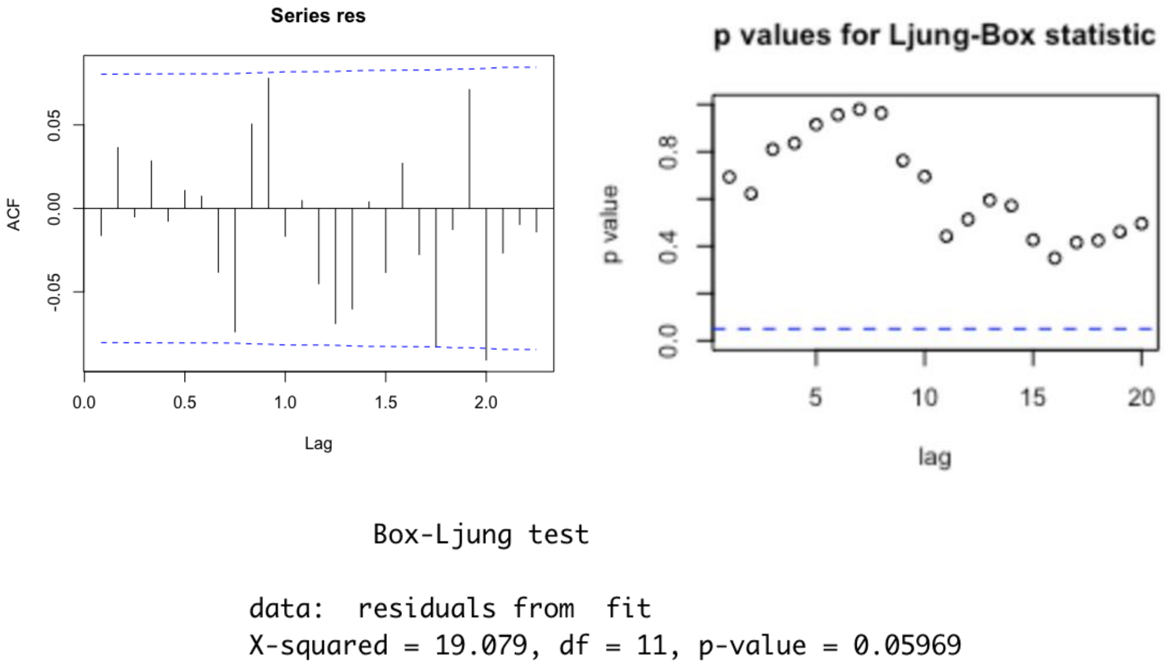
```
          Box-Ljung test

data:  residuals from  fit
X-squared = 70.884, df = 17, p-value = 1.518e-08
```

Figure 12: ACF plot, Ljung-Box test p-values plot and standard Ljung-Box test for residuals of SARIMA(0,1,2) x (1,0,0)$_{12}$

Yet, the signal expressed by the residuals at lag 12 does not only appear in this model. It is a common problem in most time series models for this data set. We will come back to address this problem in the “Residual Analysis Summary” section.

4.1.2. ***SARIMA*(4,0,3) x (0,1,2)$_{12}$**

Next, we have *SARIMA*(4,0,3) x (0,1,2)$_{12}$. For this model, we first look at the autocorrelation test of the residuals. No significant signal is exhibited at lag 12, and we decided to discard the signal at lag 24 as it only slightly passes the threshold and including this information. Also, the p-value at high lags remain greater than 0.05. We can conclude that the residuals of this model have no autocorrelation.



```
                    Box-Ljung test

        data:  residuals from  fit
        X-squared = 19.079, df = 11, p-value = 0.05969
```

Figure 13: ACF plot, Ljung-Box test p-values plot and standard Ljung-Box test for residuals of SARIMA(4,0,3) x (0,1,2)$_{12}$

However, the histogram of this model is highly concentrated in the middle, and deviation from the indicator line in the Q-Q plot is more severe. We thus propose that the residuals in this model may not following a normal distribution.

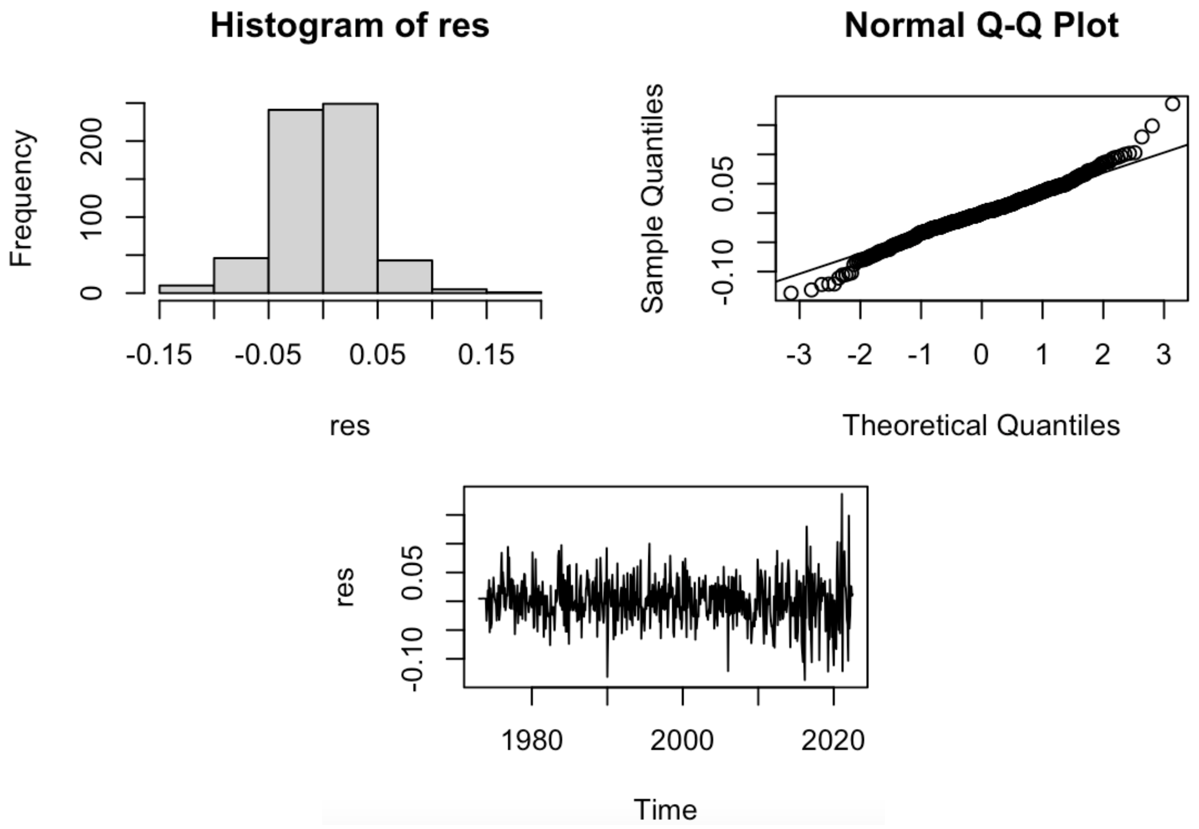
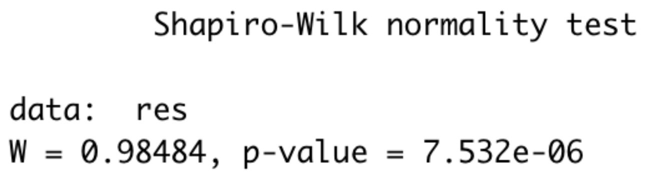


Figure 14: Histogram, Q-Q plot and time plot of residuals of SARIMA(4,0,3) x $(0,1,2)_{12}$

By performing a standard normality test, we reject the null hypothesis that the residuals are normal with a p-value of 7.532 x 10^{-6}.



```
          Shapiro-Wilk normality test

data:  res
W = 0.98484, p-value = 7.532e-06
```

Figure 15: Standard normality test for residuals of of SARIMA(4,0,3) x $(0,1,2)_1$

4.1.3. **Residual Analysis Summary**

To conclude this section, we present a summary table listing the formal normality and autocorrelation tests on the residuals of the models and summarizes each of their problems.

Models	**Normality test (p-value)**	**Ljung-Box test (p-value)**	**Problem**
SARIMA(0,1,2) x (1,0,0)$_{12}$	0.06405	1.518e-08	Autocorrelation of residuals at lag 12
SARIMA(4,0,3) x (0,1,2)$_{12}$	7.532e-06	0.05969	**Non-stationarity in variance in data**

For the first model, the problem lies on a signal of residual autocorrelation at lag 12. To adjust for this, we treat the signal as either a *MA*(12) or a *SMA*(1) process. Taking the fact that we should consider an *ARMA*(m, q) model when the residuals follow *MA*(q), we propose the following models and seek for opportunities to remove the autocorrelation:

1. *SARIMA*(2,1,12) x (1,0,0)$_{12}$ treating the residuals as *MA*(12)

2. *SARIMA*(0,1,2) x (1,0,1)$_{12}$, treating the residuals as *SMA*(1)

Applying the formal normality and autocorrelation tests again, we have the following results.

Models	**Normality test (p-value)**	**Ljung-Box test (p-value)**
SARIMA(2,1,12) x (1,0,0)$_{12}$	5.471e-05	0.0008513
SARIMA(0,1,2) x (1,0,1)$_{12}$	1.281e-06	0.0001082

In short, a mitigation of residual autocorrelations is shown, but not enough to achieve a p-value of beyond 0.05 to conclude that there are no autocorrelations. If we look at the ACF plot of one of these models (see below), further autocorrelations are exhibited at lag 23 and 24 and the models start to become complex. In addition, now the models behave poorly in the normality test, without any exception. Therefore, we conclude that the autocorrelation of residuals in our current models cannot be explained thoroughly.

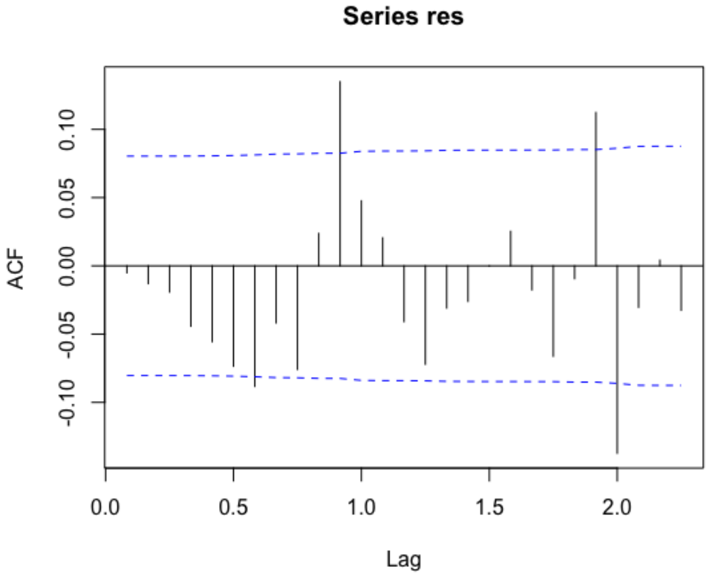
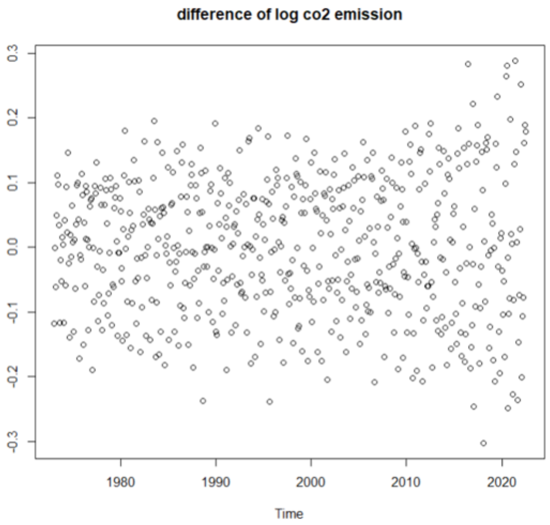


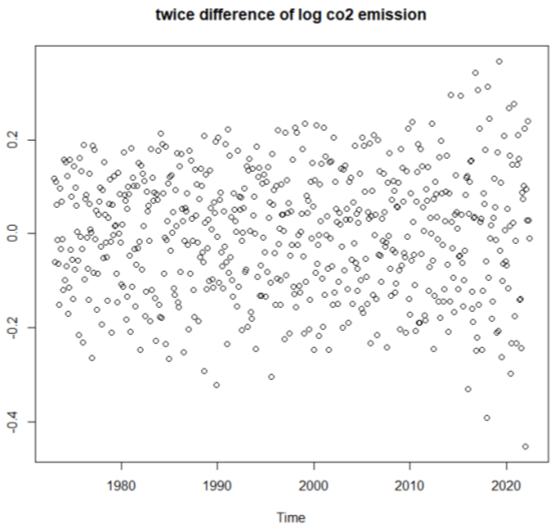
Figure 16: ACF plot for residuals of SARIMA (0,1,2) x (1,0,1)$_{12}$

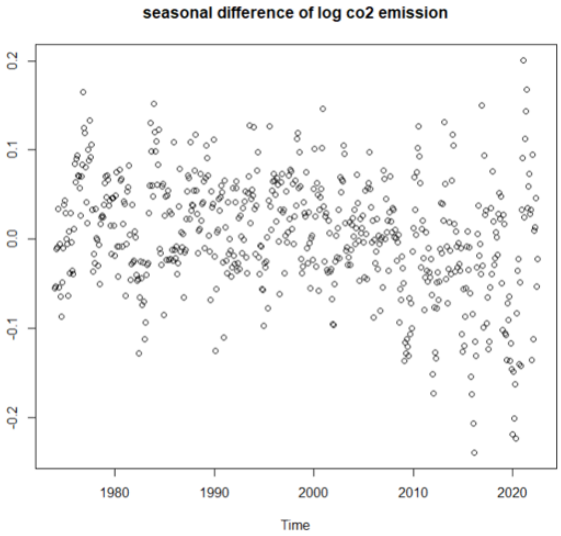
On the other hand, although the second model fails the normality test, we are still in favour of it in real-life practices, as it is possible to explain the non-normality by the unstable variance of the data (in the next sub-section), but ignoring the significant residual autocorrelation may potentially lead to overlooking some important features in forecasting future CO2 emissions.

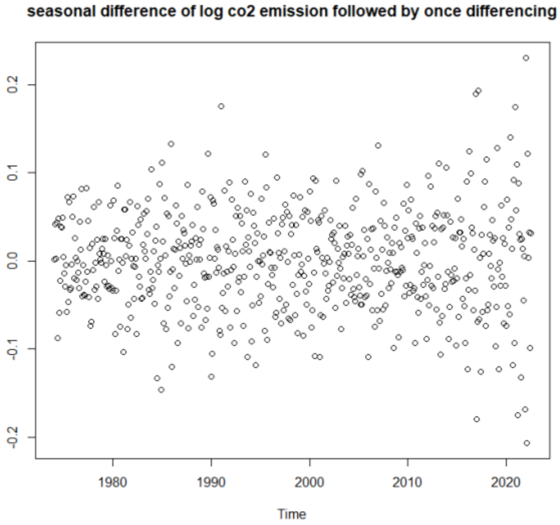
4.1.4. Non-stationarity in Variance in Data

To account for the non-normality, we go back to revise our data. We recall that despite the transformations we performed on the data, the data points exhibit a horn shape in the last few years. Clearly, the data points are unstable and the vacillation in variance in the last few years cannot be eliminated. Therefore, the data we fit our models on indeed has considerably larger variance at the end, and thus the model residuals are not normal. However, we have concluded that it is safe to assume that the data in recent years are outliers due to external social factors, so it is reasonable for our model to fail in explaining the variance of data in recent years.





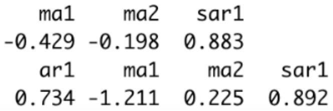
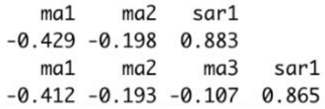
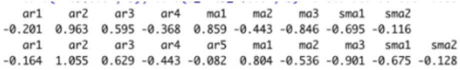
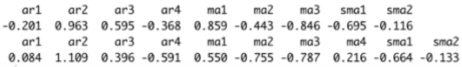




The reasons behind this non-stationarity might be the external social factors such as COVID-19 and the Russian-Ukrainian War. Since 2020 is the year when COVID-19 outbreaks, this coincides with the fact that CO2 emission starts to variate a lot from 2020 onwards, and the war between Russian and Ukrainian this year might also account for the substantial increase in CO2 emission variation as we approach the end of the timeline. These factors are out of the control of any single identity, so it is either the normality of the residuals or some potential information at high lags has to be sacrificed. Due to the fact that the residuals ACF is highly significant at lag 12 in the first models, ignoring this might constitute to loss in some useful information. This is the reason we propose that it is more reasonable to neglect the non-normality of the residuals in the last model and we prefer the last model the most.

4.2. Overfitting

To check the adequacy of our models, we have also performed over-fitting by adding 1 to the MA part and AR part respectively. Results are presented in the following table.

Model	p+1	q+1
SARIMA(0,1,2) x (1,0,0)$_{12}$	*SARIMA*(0,1,2) x (1,0,0)$_{12}$  ma1 ma2 sar1 -0.429 -0.198 0.883 ar1 ma1 ma2 sar1 0.734 -1.211 0.225 0.892 *SARIMA*(1,1,2) x (1,0,0)$_{12}$ Coefficient of a_{t-1} become too big	*SARIMA*(0,1,2) x (1,0,0)$_{12}$  ma1 ma2 sar1 -0.429 -0.198 0.883 ma1 ma2 ma3 sar1 -0.412 -0.193 -0.107 0.865 *SARIMA*(0,1,3) x (1,0,0)$_{12}$ No significant changes to existing coefficients
SARIMA(4,0,3) x (0,1,2)$_{12}$	*SARIMA*(4,0,3) x (0,1,2)$_{12}$  ar1 ar2 ar3 ar4 ma1 ma2 ma3 sma1 sma2 -0.201 0.963 0.595 -0.368 0.859 -0.443 -0.846 -0.695 -0.116 ar1 ar2 ar3 ar4 ar5 ma1 ma2 ma3 sma1 sma2 -0.164 1.055 0.629 -0.443 -0.082 0.804 -0.536 -0.901 -0.675 -0.128 *SARIMA*(5,0,3) x (0,1,2)$_{12}$ No significant changes to existing coefficients	*SARIMA*(4,0,3) x (0,1,2)$_{12}$  ar1 ar2 ar3 ar4 ma1 ma2 ma3 sma1 sma2 -0.201 0.963 0.595 -0.368 0.859 -0.443 -0.846 -0.695 -0.116 ar1 ar2 ar3 ar4 ma1 ma2 ma3 ma4 sma1 sma2 0.084 1.109 0.396 -0.591 0.550 -0.755 -0.787 0.216 -0.664 -0.133 *SARIMA*(4,0,4) x (0,1,2)$_{12}$ Coefficient of Z_{t-2} become too big

In our case, we mainly utilize two facts to discard the over-fitted models: Either the coefficient of the newly added variable is too insignificant, or the coefficient of certain variable becomes too big for a time-series process. With the idea of parsimony, we conclude that our original models are adequate.

5. Forecasting

We use the above proposed models to make prediction on last 12 months of our data. We then compare the predicted values to the original values and evaluate the performance of the two models.

		***SARIMA*(0,1,2) x (1,0,0)$_{12}$**		***SARIMA*(4,0,3) x (0,1,2)$_{12}$**	
Time	**Actual**	**Predicted**	**CI**	**Predicted**	**CI**
2021-08-01	5.1558	5.1306	[5.0343, 5.2270]	5.1253	[5.0463, 5.2044]
2021-09-01	4.9202	4.9242	[4.8111, 5.9373]	4.9018	[4.8080, 4.9957]
2021-10-01	4.7731	4.8343	[4.7156, 4.9530]	4.8029	[4.7031, 4.9028]
2021-11-01	4.6998	4.7650	[4.6410, 4.8891]	4.7390	[4.6364, 4.8417]

2021-12-01	4.7279	4.9387	**[4.8096, 5.0679]**	4.8750	**[4.7692, 4.9808]**
2022-01-01	4.9795	4.9572	[4.8232, 5.0912]	4.8939	[4.7864, 5.0014]
2022-02-01	4.7786	4.9624	**[4.8236, 5.1012]**	4.7849	[4.6749, 4.8948]
2022-03-01	4.6727	4.7620	[4.6187, 4.9057]	4.7051	[4.5936, 4.8166]
2022-04-01	4.5956	4.6910	[4.5432, 4.8388]	4.6007	[4.4871, 4.7144]
2022-05-01	4.7570	4.8045	[4.6524, 4.9566]	4.7232	[4.6080, 4.8384]
2022-06-01	4.9463	5.0583	[4.9021, 5.2146]	4.9252	[4.8081, 5.0423]
2022-07-01	5.1252	5.1883	[5.0279, 5.3486]	5.0943	[4.9756, 5.2130]

This table shows the predicted values and 95% confidence interval of each model respectively. In model 1 (*SARIMA*(0,1,2) x (1,0,0)$_{12}$), there are 2 values lying outside the confidence interval, which are 2021-12-01 and 2022-02-01. For model 2 (*SARIMA*(4,0,3) x (0,1,2)$_{12}$), only 1 value lies outside the confidence interval, which is 2021-12-01. Model 2 slightly performs better,

As we take log transformation before fitting the model, we use exponential function to return the data back to original scale. Also, we calculate the RMSE (Root Mean Squared Error) of the prediction.

Time	**Actual**	***SARIMA*(0,1,2) x (1,0,0)$_{12}$**	***SARIMA*(4,0,3) x (0,1,2)$_{12}$**
2021-08-01	173.430	169.1229	168.2327
2021-09-01	137.034	137.5832	134.5377
2021-10-01	118.284	125.7529	121.8693
2021-11-01	109.920	117.3323	114.3235
2021-12-01	113.062	139.5915	130.9700
2022-01-01	145.396	142.1950	133.4716
2022-02-01	118.939	142.9375	119.6835
2022-03-01	106.989	116.9825	110.5129
2022-04-01	99.048	108.9600	99.5563
2022-05-01	116.396	122.0584	112.5242
2022-06-01	140.657	157.3285	137.7230
2022-07-01	168.204	179.1614	163.0871
RMSE		**9.3039**	**0.0723**

The RMSE of model 1, 9.3039, is much greater than the RMSE of model 2, 0.0723. Model 2 performs better than model 1, and it can predict the values at a high accuracy. Therefore, we suggest using model 2, in which the suggestion is consistent with previous section. Finally, we plot the actual and predicted values on the same graph. The plot only shows the data after 2010 so as to have a clearer look on the prediction. The black line is the actual values, and the red line is the predicted values, with blue dashed lines indicating the confidence intervals.

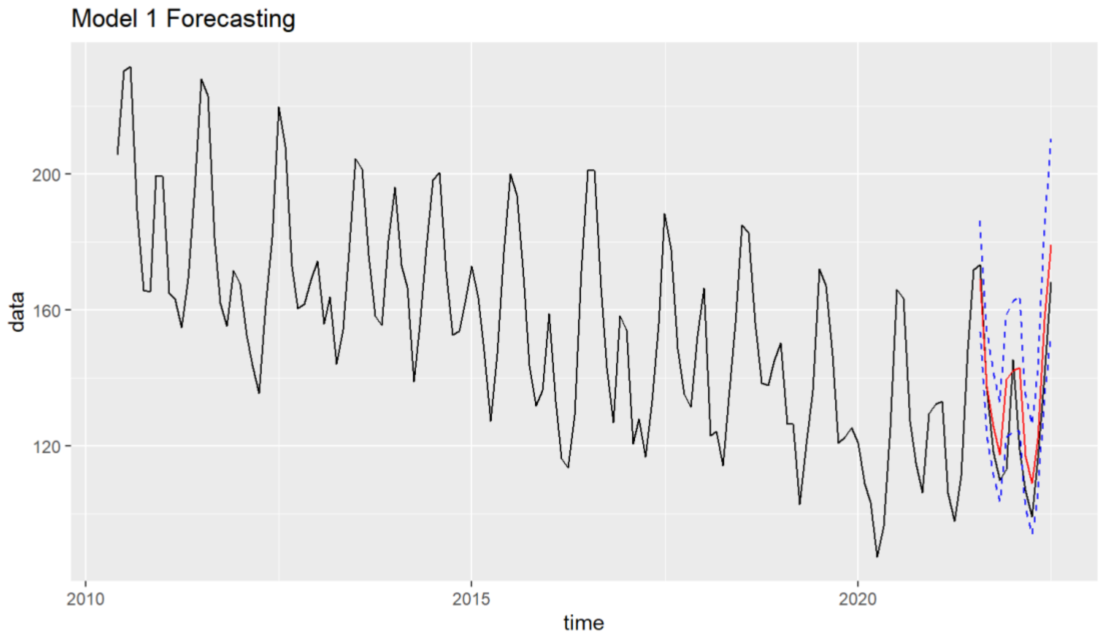


Figure 17: Plot on actual and predicted values of SARIMA(0,1,2) x (1,0,0)$_{12}$

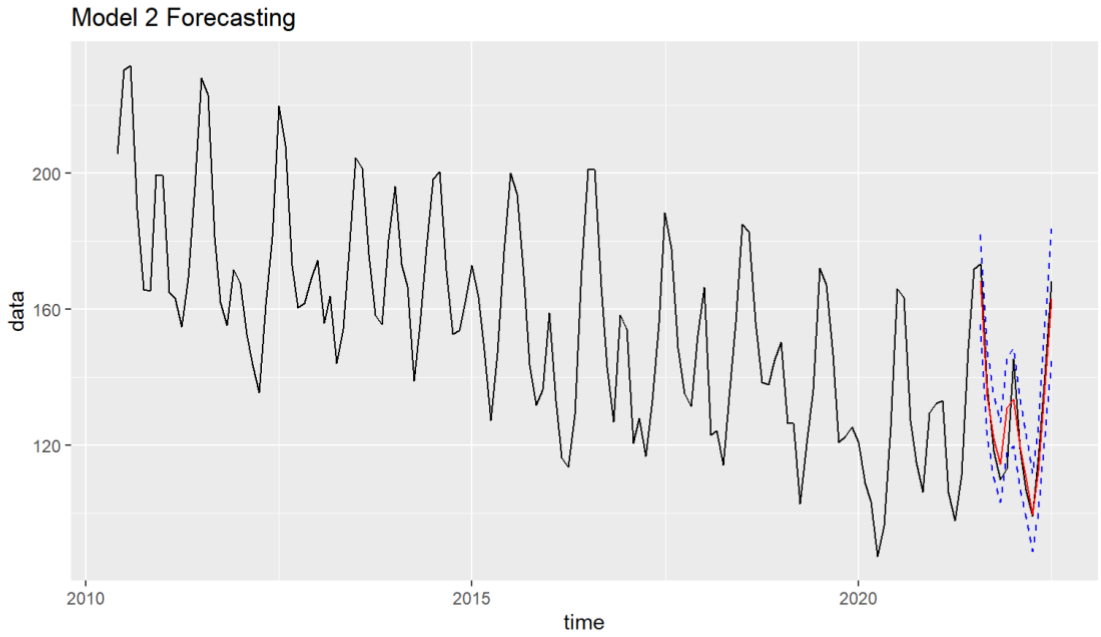


Figure 18: Plot on actual and predicted values of SARIMA(4,0,3) x (0,1,2)$_{12}$

6. Conclusion

We have obtained two models, *SARIMA*(0,1,2) x $(1,0,0)_{12}$ and *SARIMA*(4,0,3) x $(0,1,2)_{12}$. The coefficients of the models are significant. Model 1 passes Normality test while model 2 passes Ljung-Box test. The residual ACF of model 1 is still significant, hence it is not preferred. Due to the non-stationarity in variance of our data in recent years, we still accept model 2 as our final model even it does not pass the normality test for the residuals.

The prediction in last 12 values, from August 2021 to July 2022, further validates our choice. The RMSE of model 2 is relatively low, indicating that *SARIMA*(4,0,3) x $(0,1,2)_{12}$ can predict CO_2 emission fairly well.

As mentioned in above section, our data is indeed non-stationary after log transformation and seasonal differencing. There are different socio-political factors contributing to the non-stationarity. This might be the limitation of our model. However, the non-stationarity is only present in the recent 5 years so this is not an issue in general. Despite the great variance in these last few years, *SARIMA*(4,0,3) x $(0,1,2)_{12}$ is the best possible model we can obtain.

7. Appendix

The code for this project was written in the R Programming Language. The repository is open-sourced on GitHub at https://github.com/Warabi1915181/STAT4601-Time-Series-Analysis

Group member	**Contribution** **(Presentation, Report, Codes)**
Emad Akhras	PPT and report formatting, introduction, stationarity tests
Cheung Ho Ting	Model selection and model fitting
Ng Chiu Fai	Residual analysis and overfitting
Cheung Hiu Tung	Prediction, forecasting, conclusion

Works Cited

Administration, U. E. (December 2021). *U.S. Energy-Related Carbon Dioxide Emissions, 2020.* Washington, DC 20585: U.S. Department of Energy.