CO2 Data Time Series Analysis

Name: Theodoros Mamalis

Abstract

This project concerns the Wisconsin CO2 dataset. It focuses on the monthly CO2 emissions of Wisconsin, and analyzes it using time-series analysis techniques. The methods that were uses were ARIMA Fitting (section Analysis B) and Spectral Analysis (secion Analysis C), to fit and verify the models. Two SARIMA models are proposed and the best is chosen according to its diagnostic plots and significance of estimated parameters. Moreover, in section Analysis B, the next five future CO2 emission values are predicted using the best model. The findings are that an ARIMA model with seasonality of 12 months describe the model well. This is expected as CO2 emissions depend on variables that repeat themselves biyearly, e.g., industrial emission patterns, traffic patterns, or household CO2 emission patterns.

Introduction

The data concerns Carbon dioxide (CO2) in ambient and standard air samples. These samples were collected inside glass flasks. The location where the samples were taken is in Park Fall, Wisconsin, USA. The data was collected at various days within 1994-2021. The sampling interval was approximately weekly for fixed sites, which is the case for this dataset. The dataset contains a handful of parameters including the datetime the sample was collected, the value of CO2 in the sample in micromol/mol (parts per million (ppm) or 10-6 mol CO2 per mol of dry air), uncertainties of measurements, location and height where the sample was taken, whether a sample was rejected after the collection process or not.This dataset was downloaded from https://www.esrl.noaa.gov/gmd/dv/data/ , from the "Greenhouse Gases" category. In specific, the dataset used in this project can be found in this link. This dataset will be analyzed using time-series analysis methods in section Analysis B, and among two proposed ARIMA model, the best will be chosen. This model will be used to predict the CO2 emission values for Wisconsin for the future five points, i.e., for the next five months. Lastly, spectral analysis methods will be used in section Analysis C to analyze the dataset, and also verify the results of part B.

Statistical Methods

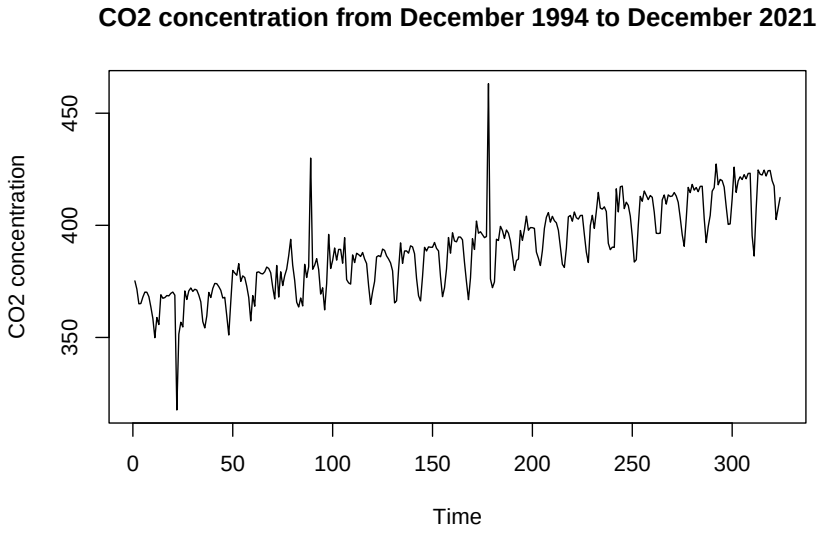
The methods that will be used will be time-series analysis methods and spectral analysis methods. The former concerns analyzing time-series such as the Wisconsin dataset, and this analysis will be made in section Analysis B that follows. The analysis includes splitting the data into train and testing datasets, and using the train dataset to fit two models. Then, the performance of these models will be used on the test dataset. The best of these models will be used to predict the CO2 emissions for the five time points, that is, for the next five months.

Moreover, after plotting the time-series in question, there seems to be seasonality existing in the CO2 emissions versus month as the timescale. For this reason, in section Analysis C, Spectral Analysis will be used, which is suited for finding underlying periodicity patterns.

Analysis B

a) This section will address the question of fitting an appropriate SARIMA model to the CO2 concentration, using month as the timescale, on training data, and checking its accuracy on testing data.

The time series plot of the concentration of CO2 is:



There is a clear increasing trend in the data. The variance is not increasing but it is mostly the same. A clear periodical (seasonal) pattern can be observed. The data is not stationary since there is an increasing trend. Two to three large spikes can be observed in the data, however, since the rest of the data look ok, these three points are not expected to affect the analysis and therefore, they will be kept.

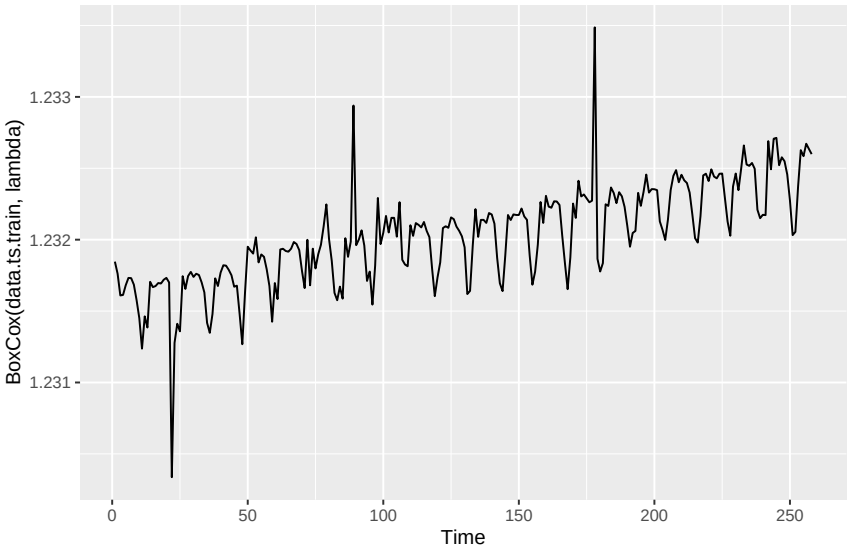
b) Then the dataset is split into a 80/20 train-test split:

At this point the dataset is split into a 80/20 train-test split. The train dataset will be used from now and until part d) of this section, where the dataset will be used to test the performance of the two models. Section Analysis C is not concerned with a train-test split but uses the original data for conducting the analysis.

Next, it will be checked whether a boxcox transformation is need:

```
## [1] -0.8049131
```

It seems that a transformation could prove useful. The resulting time series is:



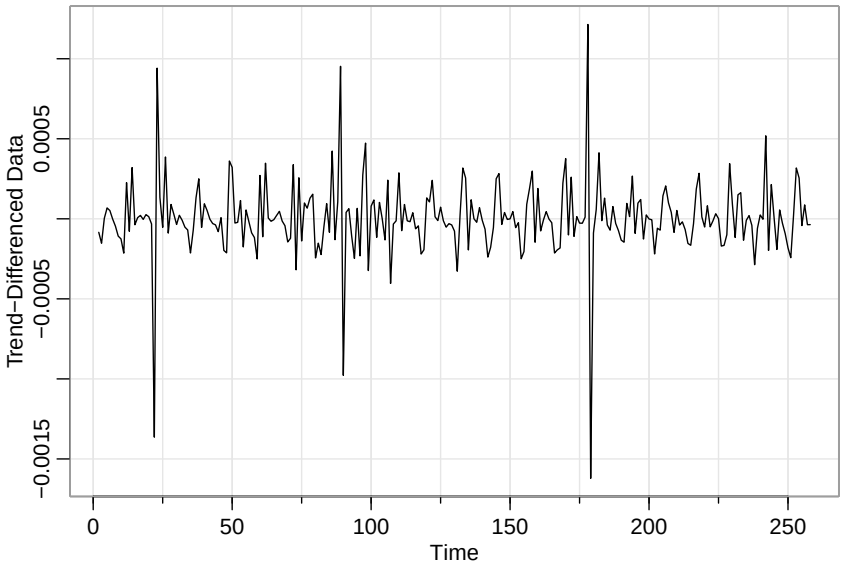
The plot looks similar to the original plot, however, the y-axis values are smaller.

The transformed data is:

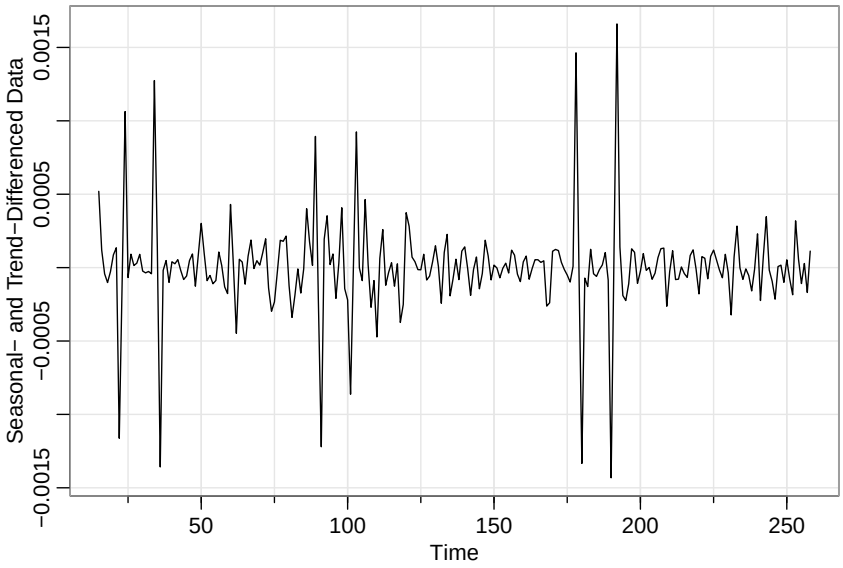
A test for differencing is made:

```
##
##  KPSS Test for Level Stationarity
##
## data:  transf.data.train
## KPSS Level = 3.648, Truncation lag parameter = 5, p-value = 0.01
```

The test suggests differencing. This was to be expected from the time series depicted in the earlier plot. By differencing, the data does not seem stationary since the variance is not constant enough at each time point:

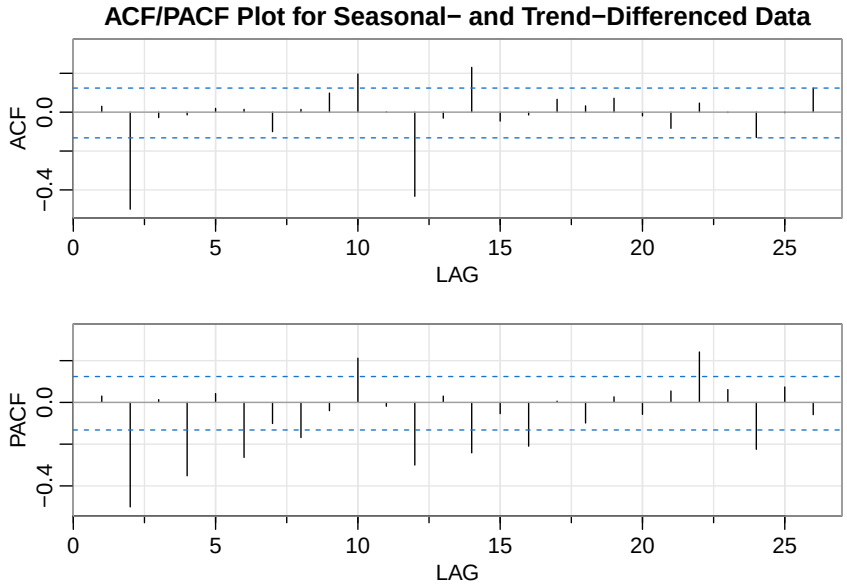


By seasonal differencing with $s = 12$ the residuals seems to be more constant, and the mean seems close to zero:



Therefore, the result looks mostly like a white noise except the 6 spikes that occured, because of the spikes in the original dataset.

C.i) The ACF and PACF plots are:



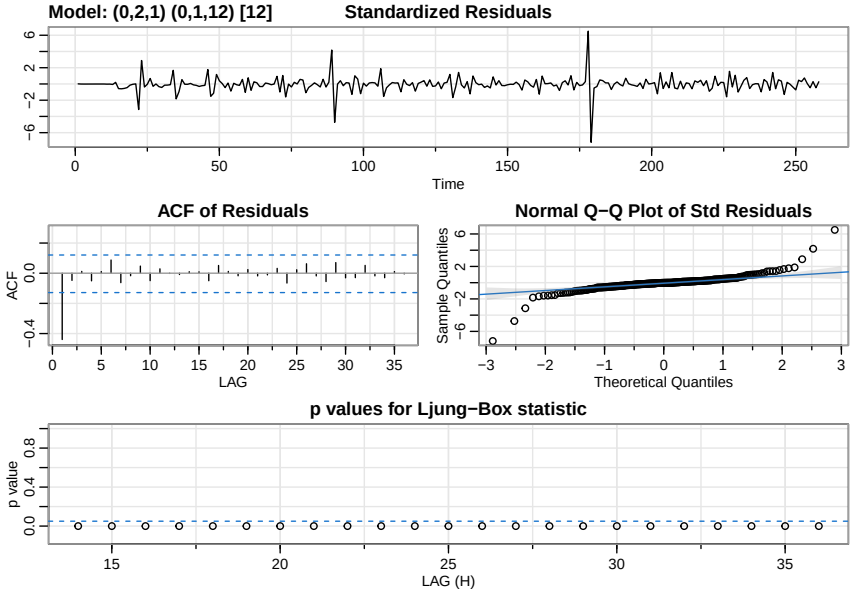
Therefore, concerning the seasonal part: The ACF is cutting of after 12 lags, and the PACF is tailing off. Therefore, $Q = 12$ and $P = 0$. Then, concerning the nonseasonal: The ACF cuts off after 1 tick and the PACF tails off. Therefore $p = 0$ and $q = 1$.

Therefore, the model is ARIMA(0,2,1)x(0,1,12)(12), where $d = 1, D = 1$ because of the two-times differencing and one-time seasonal-differencing, respectively.

The diagnostic plots are:

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##           ma1     sma1     sma2     sma3    sma4     sma5    sma6     sma7
##       -0.9999  -0.8930  -0.0516  -0.0246  0.2633  -0.3643  0.2822  -0.2310
## s.e.   0.0140   0.1703   0.1964   0.1028  0.2418   0.3018  0.2618   0.1667
##          sma8   sma9    sma10   sma11    sma12  constant
##       -0.0421  0.312  -0.1872  0.1954  -0.2555    2.2692
## s.e.   0.1571  0.173   0.1767  0.1960   0.1536  332.2793
##
## sigma^2 estimated as 90.96:  log likelihood = -931.39,  aic = 1892.78
##
## $degrees_of_freedom
## [1] 230
##
## $ttable
##          Estimate       SE  t.value p.value
## ma1       -0.9999   0.0140 -71.4316  0.0000
## sma1      -0.8930   0.1703  -5.2427  0.0000
## sma2      -0.0516   0.1964  -0.2627  0.7930
```

```
## sma3      -0.0246   0.1028  -0.2393  0.8111
## sma4       0.2633   0.2418   1.0887  0.2774
## sma5      -0.3643   0.3018  -1.2071  0.2286
## sma6       0.2822   0.2618   1.0781  0.2821
## sma7      -0.2310   0.1667  -1.3860  0.1671
## sma8      -0.0421   0.1571  -0.2678  0.7891
## sma9       0.3120   0.1730   1.8034  0.0726
## sma10     -0.1872   0.1767  -1.0595  0.2905
## sma11      0.1954   0.1960   0.9968  0.3199
## sma12     -0.2555   0.1536  -1.6635  0.0976
## constant   2.2692 332.2793   0.0068  0.9946
##
## $AIC
## [1] 7.757275
##
## $AICc
## [1] 7.764791
##
## $BIC
## [1] 7.972264
```

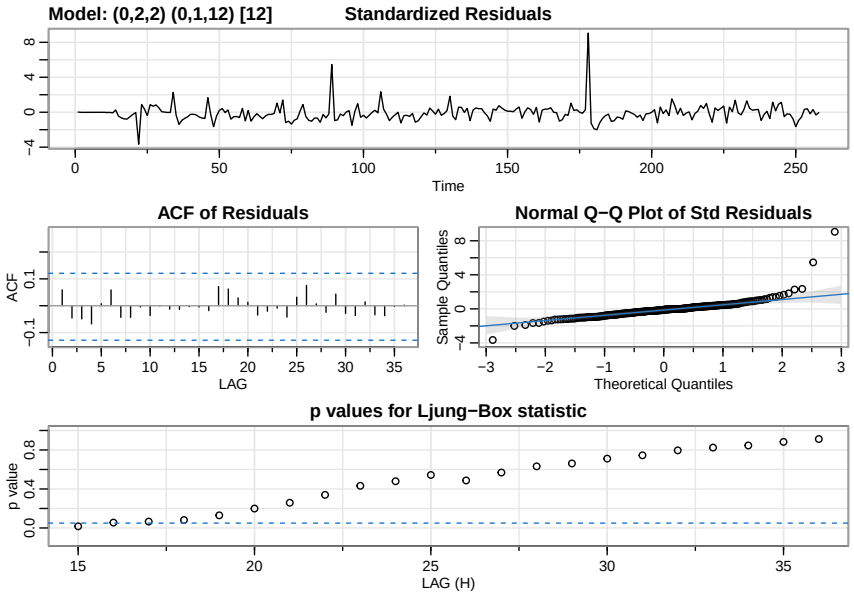


The ACF residual plot has 95 percent of the spikes within the blue band. But, the data points are not tightly gathered around the blue line. Finally, all p-values are below 0.05 in the Ljung-Box statistic plot. Therefore, this model is not a good model as far as diagnostics are concerned.

However, a good diagnostic-wise will be needed, so that significance of the parameters can be gauged Therefore, a similar model to the previous one is fit, but with $q = 2$ instead of $q = 1$. The diagnostics for the model ARIMA(0,2,2)x(0,1,12)(12) are:

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##         REPORT = 1, reltol = tol))
##
```

```
## Coefficients:
##           ma1     ma2     sma1     sma2    sma3    sma4     sma5    sma6
##       -1.9157  0.9234  -0.9243  -0.0247  0.0219  0.1726  -0.3038  0.2373
## s.e.   0.0250  0.0261   0.1472   0.1649  0.1050  0.2213   0.2804  0.2447
##          sma7     sma8    sma9   sma10   sma11    sma12  constant
##       -0.2407  -0.0265  0.3180  -0.152  0.1949  -0.2690    2.2675
## s.e.   0.1686   0.1415  0.1929   0.212  0.2335   0.1752  253.4310
##
## sigma^2 estimated as 53.23:  log likelihood = -865.2,  aic = 1762.4
##
## $degrees_of_freedom
## [1] 229
##
## $ttable
##          Estimate       SE  t.value p.value
## ma1       -1.9157   0.0250 -76.5945  0.0000
## ma2        0.9234   0.0261  35.3377  0.0000
## sma1      -0.9243   0.1472  -6.2811  0.0000
## sma2      -0.0247   0.1649  -0.1497  0.8812
## sma3       0.0219   0.1050   0.2083  0.8352
## sma4       0.1726   0.2213   0.7800  0.4362
## sma5      -0.3038   0.2804  -1.0835  0.2797
## sma6       0.2373   0.2447   0.9697  0.3332
## sma7      -0.2407   0.1686  -1.4279  0.1547
## sma8      -0.0265   0.1415  -0.1875  0.8515
## sma9       0.3180   0.1929   1.6485  0.1006
## sma10     -0.1520   0.2120  -0.7171  0.4741
## sma11      0.1949   0.2335   0.8348  0.4047
## sma12     -0.2690   0.1752  -1.5350  0.1262
## constant   2.2675 253.4310   0.0089  0.9929
##
## $AIC
## [1] 7.222946
##
## $AICc
## [1] 7.231575
##
## $BIC
## [1] 7.452269
```

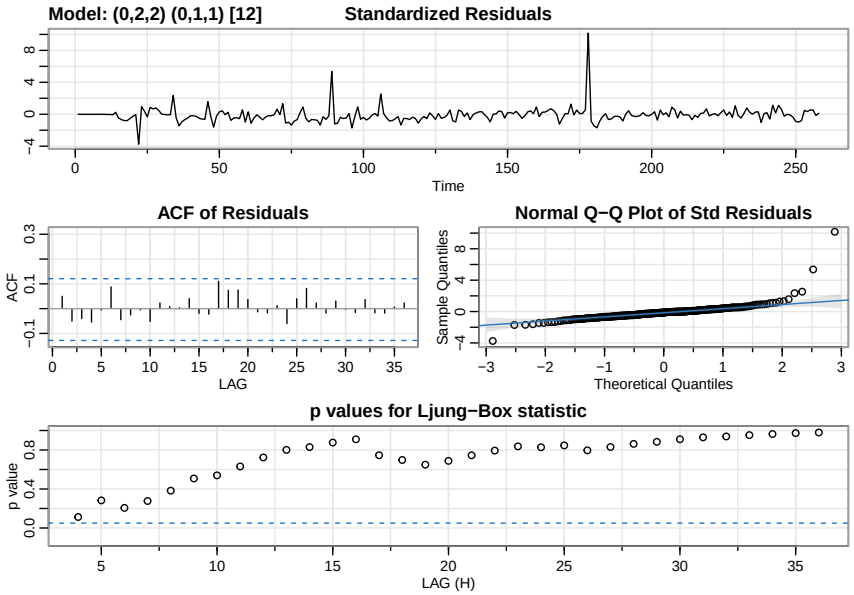


The ACF residual plot has all spikes within the blue band. Moreover, the data points are tightly gathered around the blue line in the QQ plot except a few, which are the points that represents large spikes in the time series plot presented in the beginning. Finally, all but one p-values are over 0.05 in the Ljung-Box statistic plot. Therefore, this model is a good model as far as diagnostics are concerned.

However, most variables are not significant. After trial and error, the ARIMA(0,2,2)x(0,1,1)(12) model below was found to have all variables to be significant:

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##           ma1     ma2     sma1  constant
##       -1.9375  0.9376  -0.9324    2.2693
## s.e.   0.0360  0.0343   0.0759  244.2321
##
## sigma^2 estimated as 60.96:  log likelihood = -870.99,  aic = 1751.99
##
## $degrees_of_freedom
## [1] 240
##
## $ttable
##          Estimate       SE  t.value p.value
## ma1       -1.9375   0.0360 -53.7894  0.0000
## ma2        0.9376   0.0343  27.3384  0.0000
## sma1      -0.9324   0.0759 -12.2845  0.0000
## constant   2.2693 244.2321   0.0093  0.9926
##
## $AIC
## [1] 7.180278
##
## $AICc
```

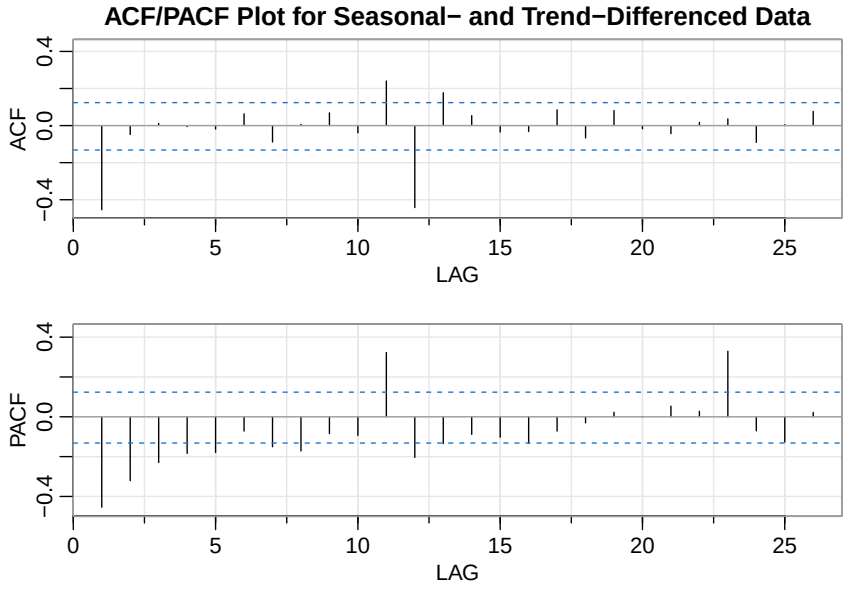
```
## [1] 7.180964
##
## $BIC
## [1] 7.251941
```



The ACF residual plot has all spikes within the blue band. Moreover, the data points are tightly gathered around the blue line in the QQ plot except 3, which are the points that represents the spikes in the data. Finally, all but p-values are over 0.05 in the Ljung-Box statistic plot. Therefore, this model is a good model as far as diagnostics are concerned.

c.ii) A different model will be presented below for $d = 1$.

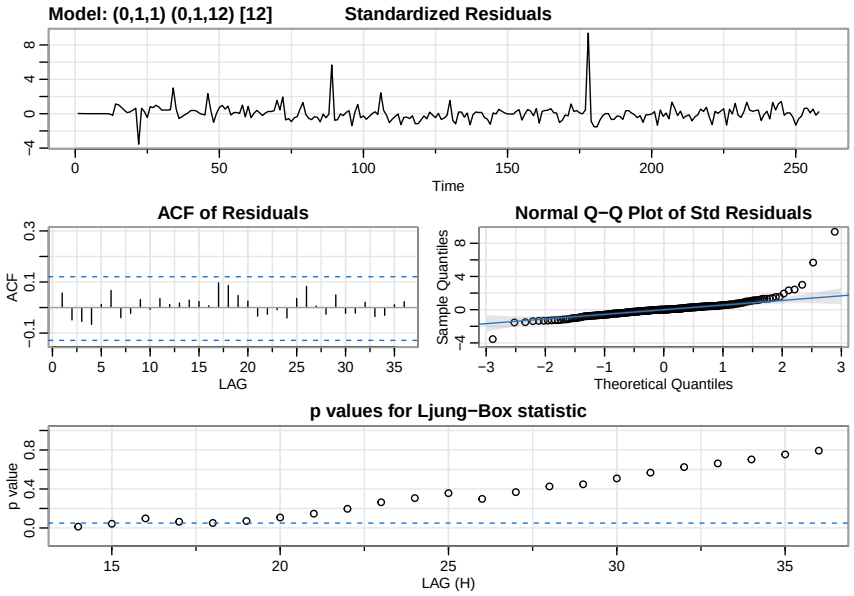
The ACF and PACF plots are:



They are similar to the **c. i)** therefore the parameters are the same with the difference that in this case $d = 1$.

The model diagnostics are:

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     include.mean = !no.constant, transform.pars = trans, fixed = fixed, optim.control = list(trace =
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##           ma1     sma1    sma2     sma3    sma4     sma5    sma6     sma7
##       -0.9467  -0.8676  -0.096  -0.0086  0.2345  -0.3658  0.3072  -0.2568
## s.e.   0.0299   0.3734   0.464   0.1008  0.4912   0.6640  0.5667   0.3191
##          sma8    sma9    sma10   sma11    sma12
##       -0.0125  0.3209  -0.2136  0.2312  -0.2727
## s.e.   0.1463  0.2587   0.2561  0.2484   0.1792
##
## sigma^2 estimated as 49.43:  log likelihood = -858.21,  aic = 1744.41
##
## $degrees_of_freedom
## [1] 232
##
## $ttable
##       Estimate     SE  t.value p.value
## ma1    -0.9467 0.0299 -31.7118  0.0000
## sma1   -0.8676 0.3734  -2.3233  0.0210
## sma2   -0.0960 0.4640  -0.2068  0.8363
## sma3   -0.0086 0.1008  -0.0857  0.9318
## sma4    0.2345 0.4912   0.4774  0.6335
## sma5   -0.3658 0.6640  -0.5509  0.5823
## sma6    0.3072 0.5667   0.5421  0.5883
## sma7   -0.2568 0.3191  -0.8048  0.4218
## sma8   -0.0125 0.1463  -0.0857  0.9318
## sma9    0.3209 0.2587   1.2403  0.2161
## sma10  -0.2136 0.2561  -0.8340  0.4051
## sma11   0.2312 0.2484   0.9306  0.3530
## sma12  -0.2727 0.1792  -1.5215  0.1295
##
## $AIC
## [1] 7.120046
##
## $AICc
## [1] 7.126478
##
## $BIC
## [1] 7.320118
```

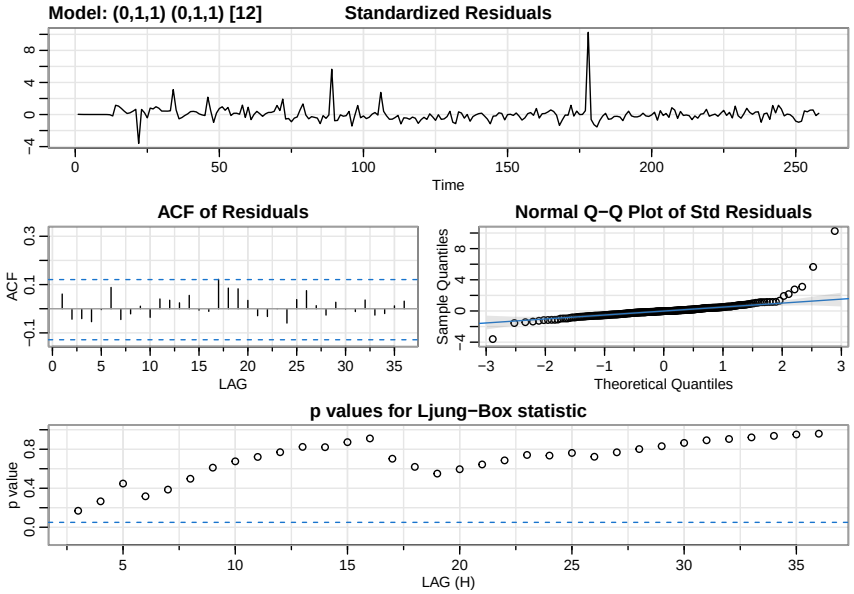


The ACF residual plot has all spikes within the blue band. Moreover, the data points are mostly tightly gathered around the blue line in the QQ plot except a few, which are the points that represents the spikes in the data. Finally, all but p-values are over 0.05 in the Ljung-Box statistic plot. Therefore, this model is a good model as far as diagnostics are concerned.

However, most variables are not significant. After trial and error, the ARIMA(0,1,1)x(0,1,1)(12) model below was found to have all variables to be significant. The model diagnostics are:

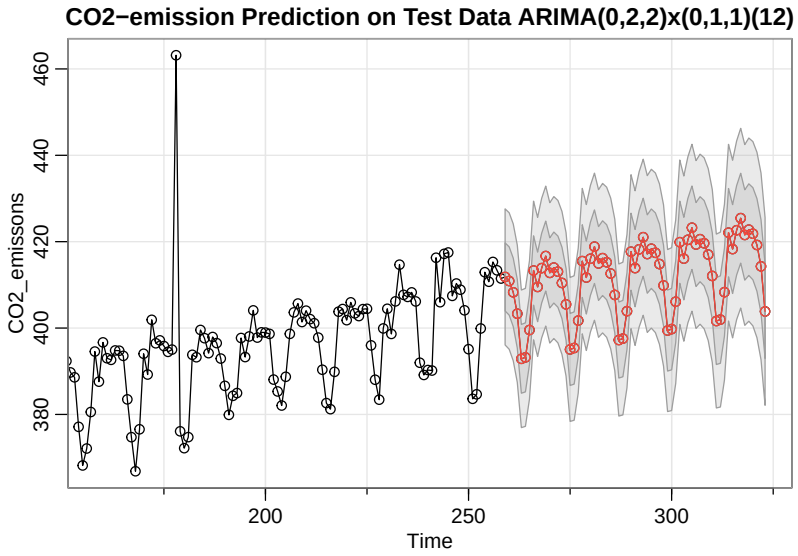
```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     include.mean = !no.constant, transform.pars = trans, fixed = fixed, optim.control = list(trace =
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##           ma1     sma1
##       -0.9519  -0.9445
## s.e.   0.0272   0.0862
##
## sigma^2 estimated as 59.99:  log likelihood = -863.88,  aic = 1733.77
##
## $degrees_of_freedom
## [1] 243
##
## $ttable
##      Estimate     SE  t.value p.value
## ma1   -0.9519 0.0272 -35.0463       0
## sma1  -0.9445 0.0862 -10.9540       0
##
## $AIC
## [1] 7.076596
##
## $AICc
## [1] 7.076799
##
```

```
## $BIC
## [1] 7.119469
```

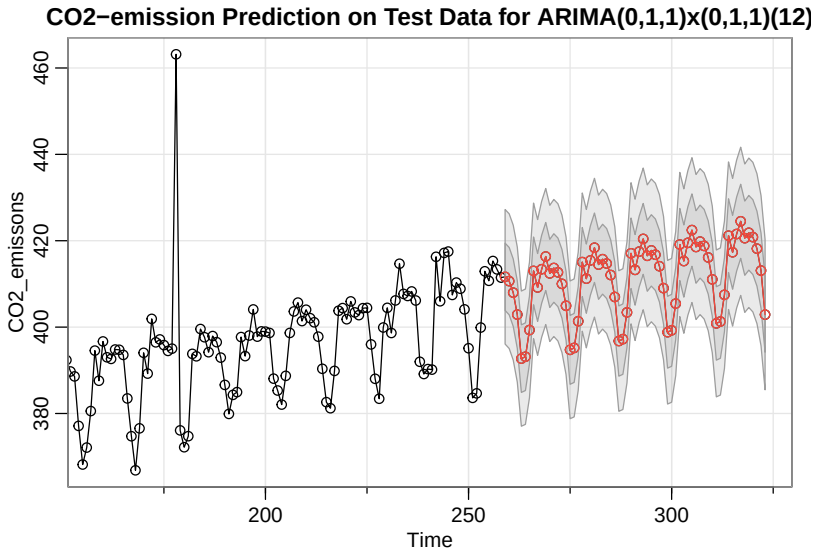


The ACF residual plot has all spikes within the blue band. Moreover, the data points are tightly gathered around the blue line in the QQ plot except a few, which are the points that represents the spikes in the data. Finally, all but p-values are over 0.05 in the Ljung-Box statistic plot. Therefore, this model is a good model as far as diagnostics are concerned.

d) The forecasts of the models are:



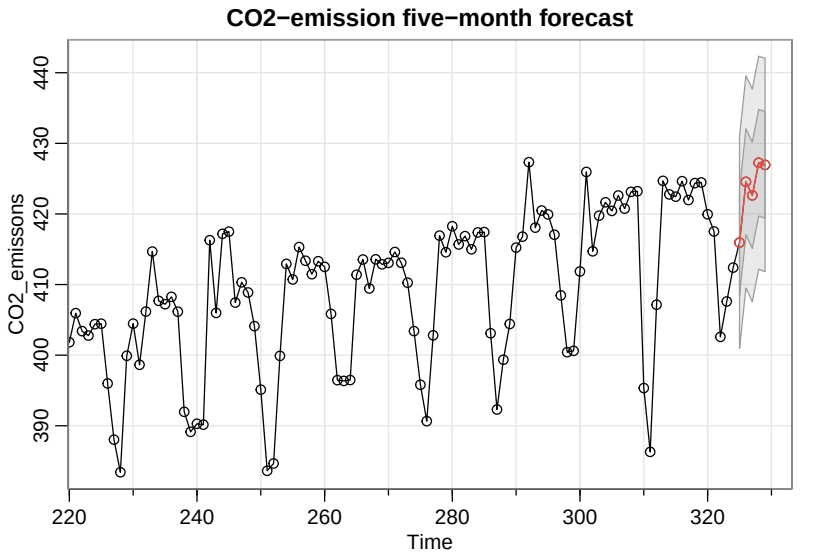

```
##                      ME     RMSE      MAE       MPE      MAPE      ACF1 Theil's U
## Test set 0.5699025 4.919397 3.524044 0.1230266 0.8622647 0.2868462 0.6139284
```




```
##                    ME     RMSE      MAE       MPE      MAPE      ACF1 Theil's U
## Test set 1.184978 5.061624 3.724484 0.2713856 0.9091125 0.2932393 0.6292779
```

According to all forecasting metrics, the ARIMA(1,2,1)x(0,1,1)(12) model is better than the ARIMA(1,1,1)x(0,1,1)(12). Therefore, the former model is chosen as the best model among the two.

e) The forecast of the future 5 values according to the best ARIMA(1,2,1)x(0,1,1)(12) model is:




```
## $pred
## Time Series:
## Start = 325
## End = 329
## Frequency = 1
## [1] 415.9423 424.5452 422.6404 427.2478 426.9476
##
## $se
## Time Series:
## Start = 325
```

```
## End = 329
## Frequency = 1
## [1] 7.502336 7.516373 7.530576 7.544942 7.559473
```

Analysis C

a) Literature review

Historically, spectral analysis has been coupled with time series analysis from as early as fifty years ago. For example, in Parzen (1967) the authors try to couple time series with spectral analysis, and discusses several foundational spectral-analysis concepts, such as the periodogram. The paper then uses those concepts to study an empirical time series. Then, a few decades later, it seems that spectral analysis, as well as time-series analysis, has not yet been incorporated by fields that, today, make heavy use of time-series analysis tools, such as Econometrics. For example, in Granger and Watson (1984), the authors are critical of econometricians that refuse to accept time-series analysis as a valid analysis tools for econometric time-series. In order to address any concerns related to the use of time-series analysis with economics data, the authors in Granger and Watson (1984) take advantage of the existence of periodicity in econometrics, and couple it with spectral analysis tools from the time-series analysis field, which can be used to analyzed periodic time-series. More attempts are made to couple spectral analysis with real-life questions in Gardner (1986). There, the concept of spectral correlation theory is discussed in the context of cyclostationary time-series. After the author introduces relevant notation, definitions and theorems, he goes on to list applications that may benefit from such a s theory. The applications include sampling and aliasing, frequency conversion, and noise in periodic circuits. Finally, spectral analysis continues to be a relevant topic, for example in Lepage and Thomson (2009), where the authors present a spectral analysis method to analyze cyclostationary time-series with applications on seismic data, in specific, recognizing the "hum" sound introduced by subtle seismic activity during the collection of various time-series data.

[1] Parzen, E. (1967). The Role of Spectral Analysis in Time Series Analysis. *Review of the International Statistical Institute*. JSTOR. https://doi.org/10.2307/1401395.

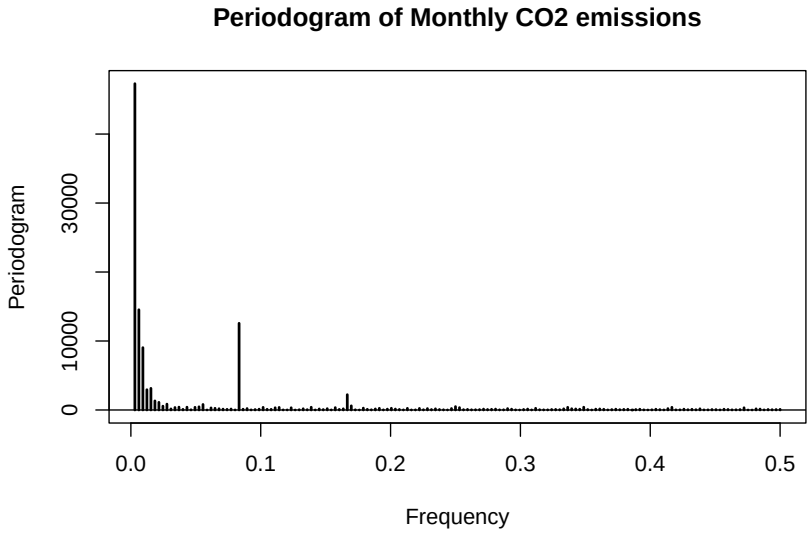
[2] Granger, C.W.J., and Watson, Mark W. (1984). Chapter 17 Time series and spectral methods in econometrics. *Handbook of Econometrics*, 979-1022. Elsevier. https://doi.org/10.1016/S1573-4412(84)02009-2.

[3] Gardner, W A. (1986). The spectral correlation theory of cyclostationary time-series. *Signal Processing*, 13–36. https://doi.org/10.1016/0165-1684(86)90092-7

[4] Lepage, K. Q., and D. J. Thomson. (2009). Spectral analysis of cyclostationary time-series: a robust method. *Geophysical Journal International*, 1199–1212. doi: 10.1111/j.1365-246X.2009.04339.x

b) The data in Part B) will be analyzed using Spectral Analysis:

Firstly, the periodogram will be used to identify frequencies of interest:



The spikes in the periodogram (those with value greater than 5000) concern the frequencies:

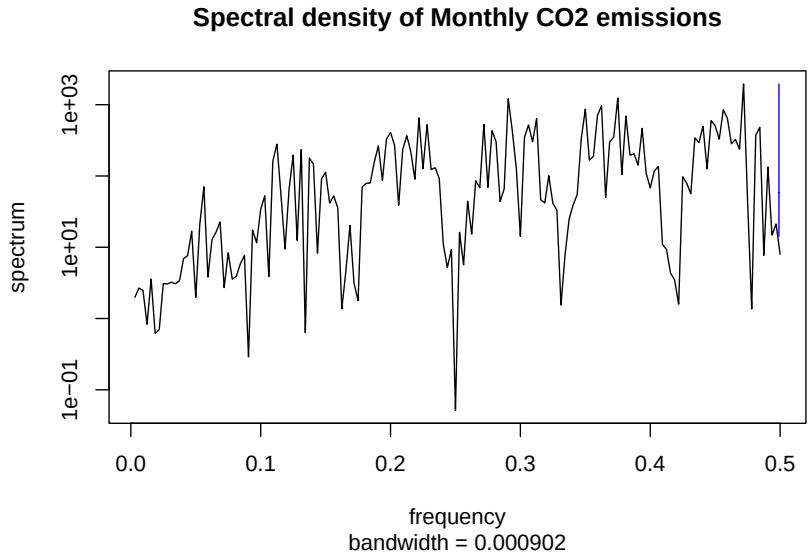
```
## [1] 0.003086420 0.006172840 0.009259259 0.083333333
```

The frequencies around 0 are to be expected, since they express the increasing trend in the data. By removing these frequencies, the last significant frequency, when converted to timescale yields a period of:

```
## [1] 12
```

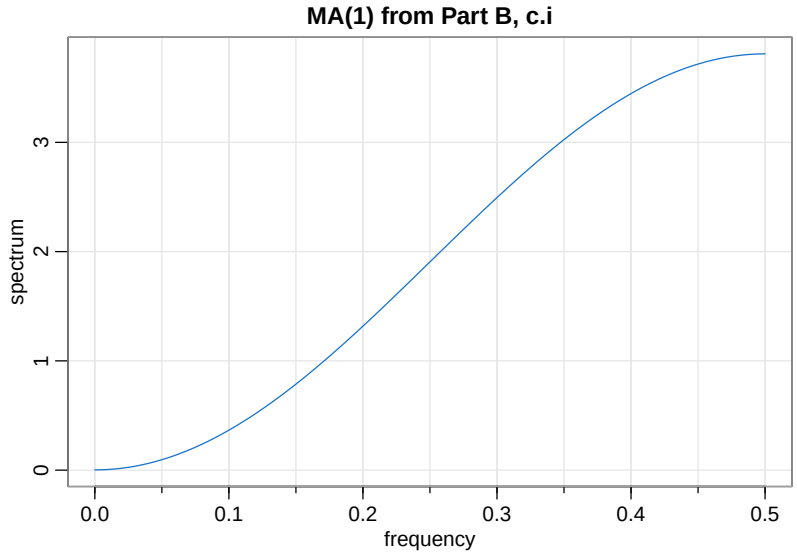
Therefore, there are repeating cycles every 12 months, that is, the cycle repeats yearly. This verifies the seasonal component of Part B) which was taken to be 12 based on the time series plots, and on the diagnostic results of the sarima() function. Moreover, this is to be expected since CO2 emissions depend on various functions of the industrial units, traffic patterns, and household CO2-emission patterns, in the state of Wisconsin, all of which more-or-less repeat themselves each year.

Then, the spectral density of the data after performing once-trend- and seasonal-differencing is calculated:



The spectral density seems to be increasing. This agrees with the ARMA(0,1,1)x(0,1,1)(12) which after once-

trend- and seasonal-differencing can be though of as an MA(1) model with the first parameter being -0.9519 according to Part B), c.i. This can be seen below by using the arma.spec function for the aforementioned MA(1) parameter:



Results and Discussion

The Analysis B section picked the ARIMA(0,2,2)x(0,1,1)(12) model as the best model but the ARIMA(0,1,1)x(0,1,1)(12) was not far behind. Even though the former is better from a statistical point of view, the latter is perhaps closer to what someone would expect from looking at the time-series plot in Analysis B. This is because, there seems to be an increasing linear trend, which is made stationary by differencing once, that is, by using $d = 1$. However, the plot could hide that the trend is in fact quadratic, hence the twice-trend-differenced model. The seasonality is 12 for both.

The Analysis C section confirmed the existence of a trend with the existence of the 0 frequency in the periodogram. It also verified the existence of the 12-month seasonality of the CO2 emission time-series, with a strong frequency at 0.083333333, which significes a period of the inverse of the frequency, that is, a CO2-emission period of 12 months.