Analysis of Air Quality

Project of Time Series Analysis

Yi Zheng 2020012859

2022/5/24

Contents

1 Introduction and Background

A substantial part of China is experiencing air pollution with severe fine particulate matter (PM) concentration and PM2.5 in particular, which refers to the fine PM with aerodynamic diameter of less than 2.5 μm. The north China Plain (NCP) that surrounds Beijing endures the most severe air pollution in the country with excessive PM2.5 concentration. In an attempt to clear up the smog, China's State Council has set a 25% PM2.5 reduction target for the NCP by 2017 relative to the 2012 level, and a specific target of no more than 60 $\mu g \cdot m^{-3}$ for Beijing's annual average.

1.1 Relative Research

Zhang et al. (2017) conducted statistical analyses on the PM2.5 data of the past 4 years from Beijing's 36 monitoring sites in conjunction with 7 years' meteorological records at 15 stations. They wanted to provide meaning and insight to the official statistics and a broader understanding of the air pollution situation in Beijing. To this end, they considered:

- two types of years, the calendar year and the seasonal year;
- two types of monitoring sites, the 11 Guokong sites and more sites in central Beijing to provide wider spatial coverage;
- two types of averages: the simple average and an adjusted average constructed under a standardized baseline meteorological condition.

Having these three perspectives in the analyses leads to a fuller view on Beijing's PM2.5 pollution in the past 4 years and 2016 in particular. The pollutant that affects people the most is particulate matter, usually abbreviated as PM and used as a measure of air pollution. Although particles with a diameter of 10 microns or less ($\leq$PM10) can penetrate and embed deep in the lungs, the ones that are more harmful to health are those with a diameter of 2.5 microns or less ($\leq$PM2.5).

In this study, we will try to focus on `O3`, another important indicator of air quality. We will build models to give predictions, as well as exploring the relations between `O3` and other indicators.

2 Data Loading and Cleaning

2.1 Overview of the Complete Data Set

The complete data set includes hourly air pollutants data from 12 nationally-controlled air quality monitoring sites. The air quality data are from the Beijing Municipal Environmental Monitoring Center (BMEMC). The meteorological data in each air quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017. Missing data are denoted as NA.

The city of Beijing established an air pollution monitoring network in January 2013 as part of the national monitoring network. There are 36 air-quality monitoring sites in Beijing, 35 of which are BMEMC sites and one at the US Embassy in Beijing.

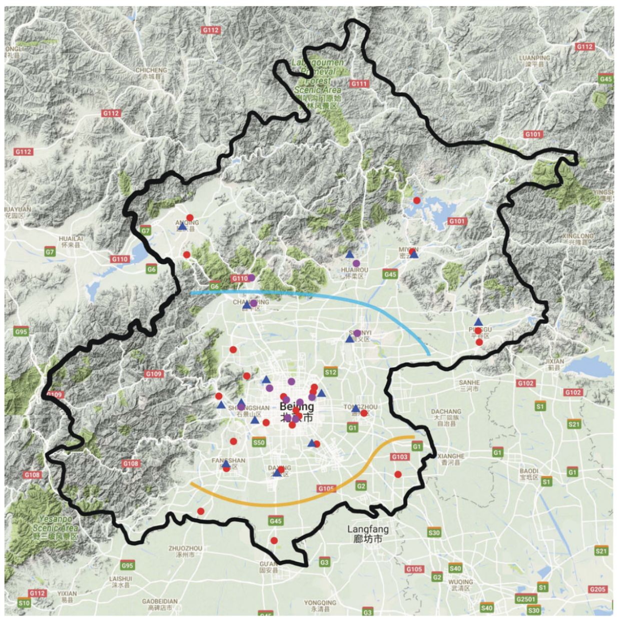


Figure 1: Location of the monitoring sites

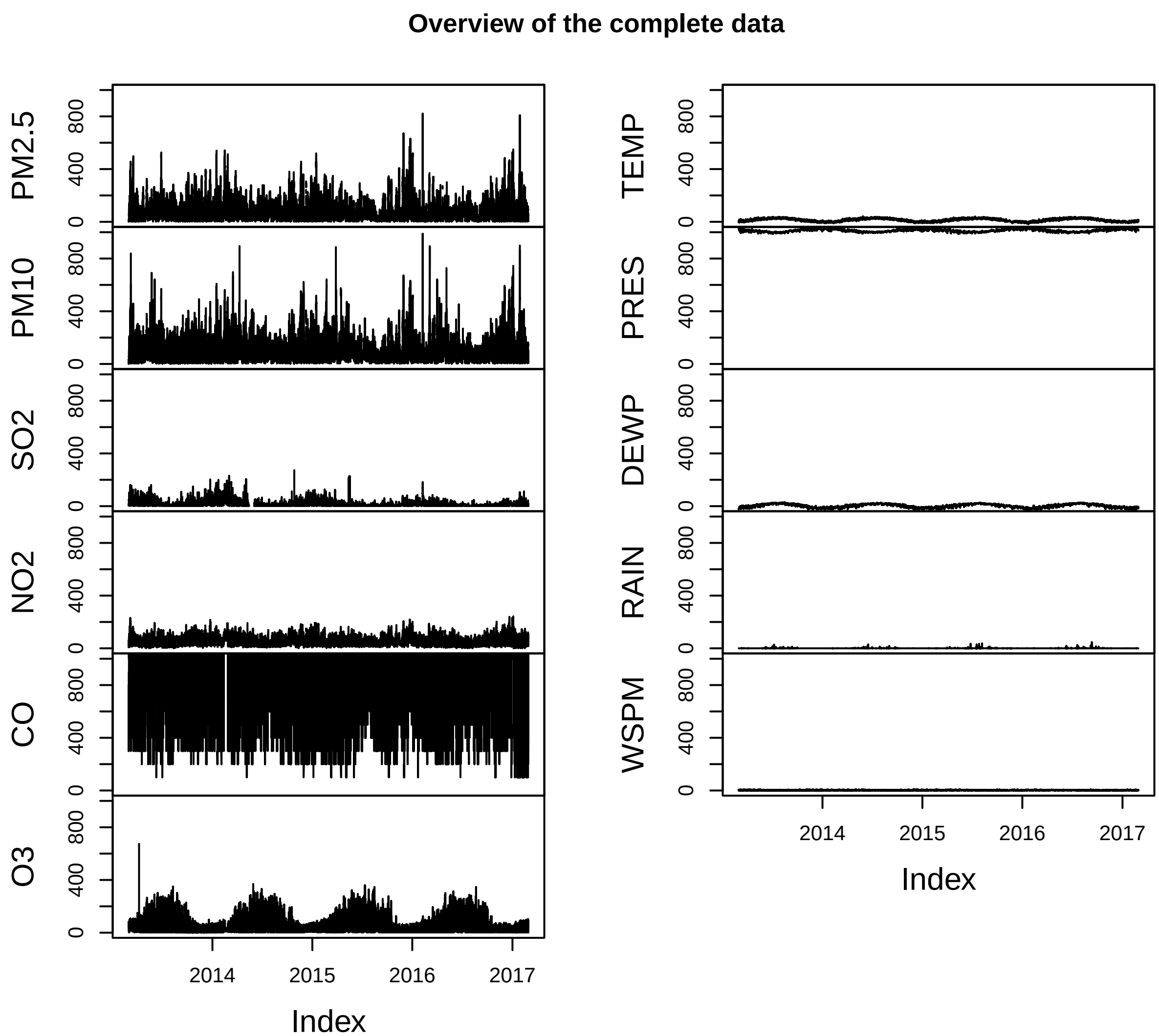
The meteorological data consist of 6 hourly observed variables: air temperature, wind direction (WD) and speed, pressure, relative humidity (or dew point temperature, DEW) and precipitation, from March 2010 to February 2017. The reason for using three more years' meteorological data is for a better construction of a spatial and temporal baseline weather condition over the study region.

As there were many missing values in January and February of 2013 in most sites when Beijing's air-quality monitoring network was first put in operation, we consider the seasonal year, namely hourly PM2.5 data ranging from March 2013 to February 2017 that makes up four seasonal years. As mentioned earlier, an advantage of using the seasonal year is that it keeps the winter season intact without breaking it into two separate years. The time unit of the study is season, which consists of segments of three months starting from March, June, September and December, which represent the four seasons of spring, summer, autumn and winter, respectively.

```
## [1] "Head of the complete data set"
```

```
## # A tibble: 6 x 18
##      No  year month   day  hour PM2.5  PM10   SO2   NO2    CO    O3  TEMP  PRES
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
## 1     1  2013     3     1     0     6     6     4     8   300    81  -0.5 1024.
## 2     2  2013     3     1     1     6    29     5     9   300    80  -0.7 1025.
## 3     3  2013     3     1     2     6     6     4    12   300    75  -1.2 1025.
## 4     4  2013     3     1     3     6     6     4    12   300    74  -1.4 1026.
## 5     5  2013     3     1     4     5     5     7    15   400    70  -1.9 1027.
## 6     6  2013     3     1     5    10    10    12    15   400    70  -2.4 1028.
## # ... with 5 more variables: DEWP <dbl>, RAIN <dbl>, wd <chr>, WSPM <dbl>,
## #   station <chr>
```



2.2 Data Preprocessing and Transformation

As a methodological example, this study only chose one of the 12 data sets to analyse, and the method applied to this data set can be conveniently transferred to other data sets.

Since we only learned how to deal with small scale of data in the class, we here transformed the data into monthly data for convenience.

```
## [1] "Head of the transformed data"
```

```
##   Group.1     PM2.5      PM10       SO2      NO2        CO       O3      TEMP
## 1 2013-03 106.22849 123.46102 37.718063 64.04525 1558.4274 63.54890  6.256989
## 2 2013-04  60.94444  90.76667 21.263254 43.59433  984.6375 76.54095 12.632361
## 3 2013-05  80.49059 138.19355 26.998656 42.26747 1097.3562 81.72987 21.929301
## 4 2013-06 110.53889 133.73889 15.279959 48.78758 1442.6236 74.80572 23.823611
## 5 2013-07  69.30511  84.12500  6.993353 43.39455 1071.2151 77.73951 27.485215
## 6 2013-08  64.23790  82.31452  6.286671 40.77480  923.5215 80.22479 27.321102
##        PRES      DEWP        RAIN     WSPM
## 1 1014.4569 -7.030511 0.020967742 1.969624
## 2 1009.9331 -3.778056 0.012222222 2.571389
## 3 1004.7483  7.802957 0.003360215 1.997177
## 4 1001.6372 17.216250 0.097916667 1.468056
## 5  997.4539 20.801210 0.263978495 1.475403
## 6 1000.7148 20.054570 0.072580645 1.579704
```

```
## [1] "English_United States.1252"
```

The data frame after transformation is 48×12.

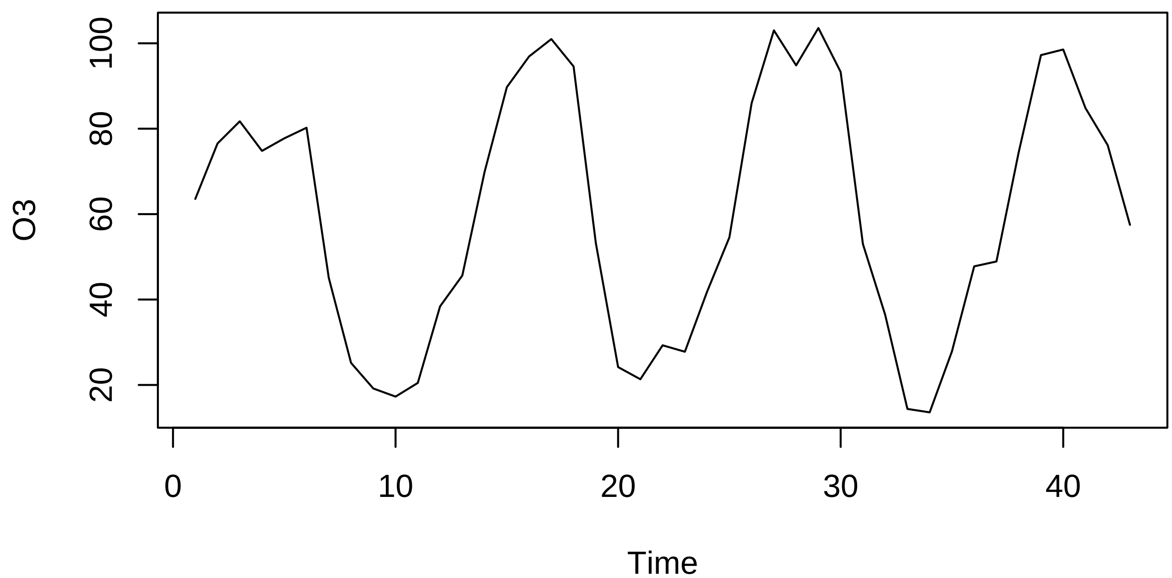
3 Data Analysis and Model Selection

In this part, we only use the first 43 data of sequence as our training data to build a model and the last five data as validation for our prediction.

3.1 ARIMA Model for `O3` Data

3.1.1 Stationarity Test

First we may plot the data sequence.



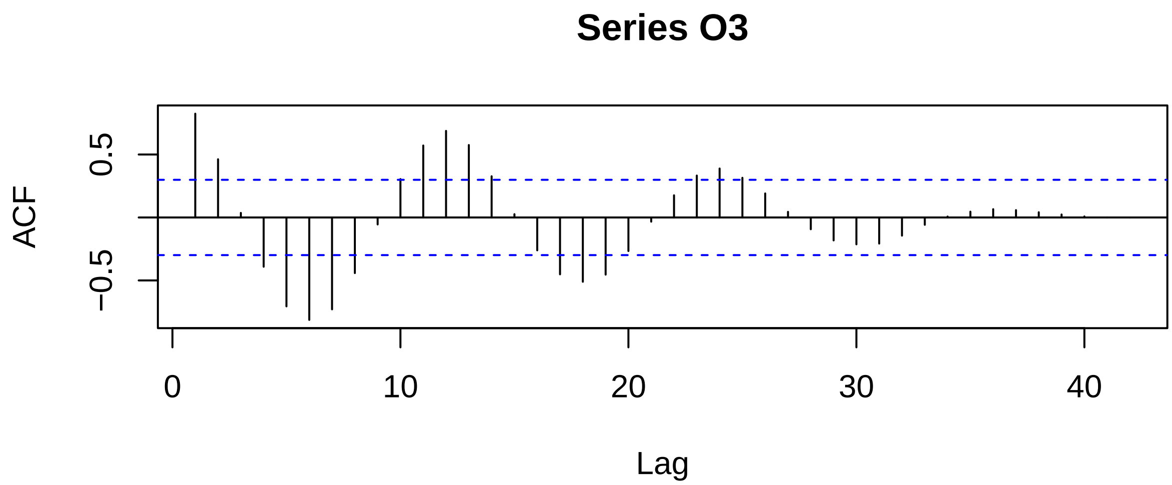
By doing ADF test, we may get the results below.

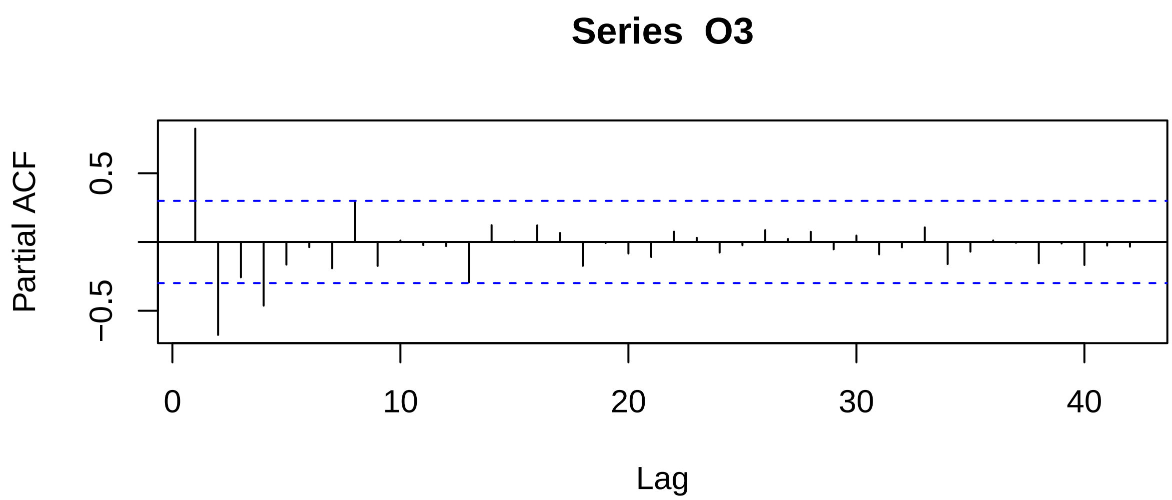
```
##
##  Augmented Dickey-Fuller Test
##
## data:  O3
## Dickey-Fuller = -6.6054, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

Therefore, it is safe to reject the null hypothesis and confirm that the sequence is stationary, then we can just build a model based on the original data without any transformation.

3.1.2 Specification

The sample ACF, PACF and EACF are plotted as below.





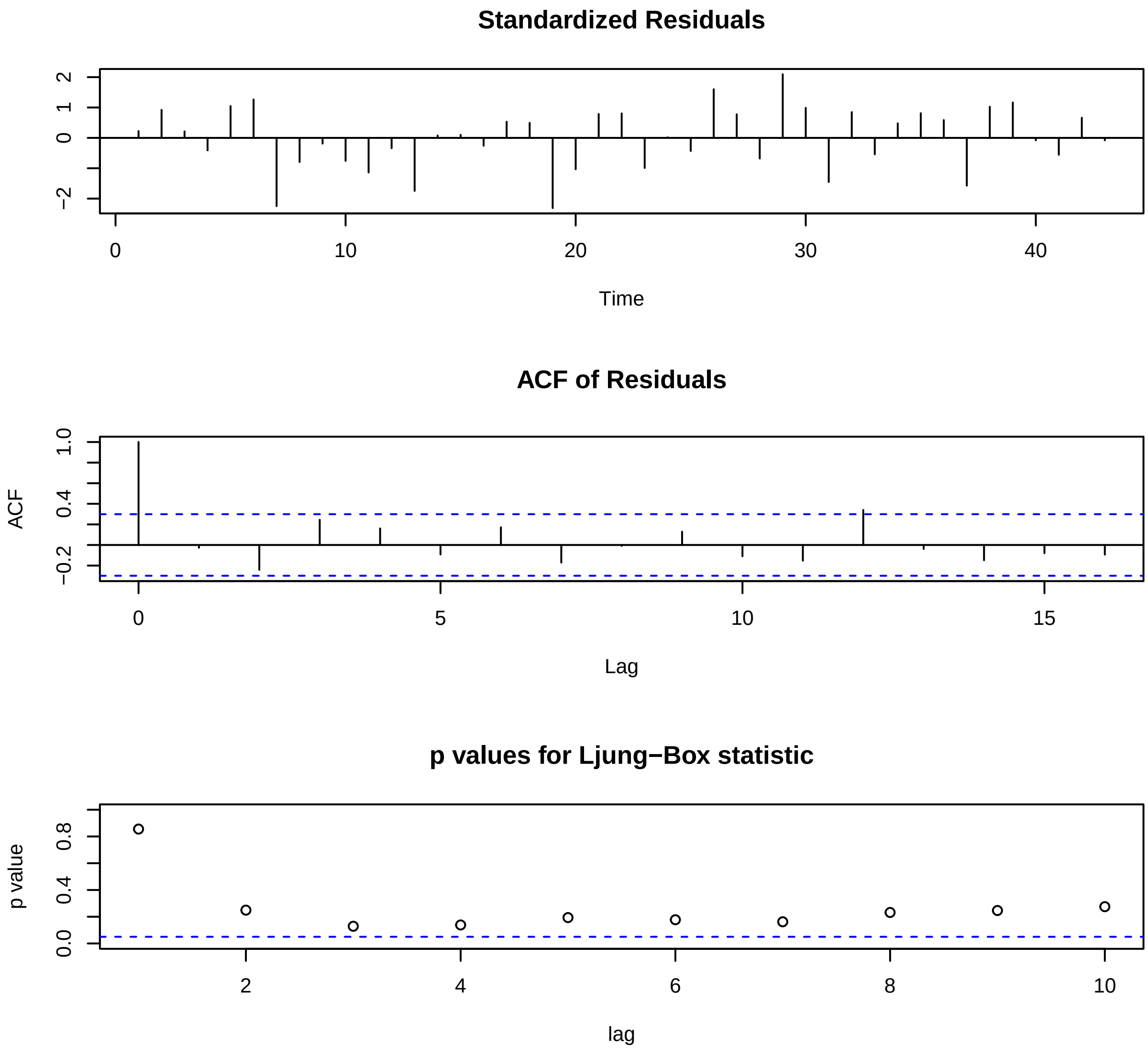

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x o x x x x x o o x  x  x  o
## 1 x x o x x x x x o o x  x  x  o
## 2 x o o o o o o o o o o  x  o  o
## 3 x x o o o o o o o o o  x  o  o
## 4 x o x o o o o o o o o  o  o  o
## 5 x o o o o o o o o o o  o  o  o
## 6 x o o x o o o o o o o  o  o  o
## 7 x o o x o o o o o o o  o  o  o
```

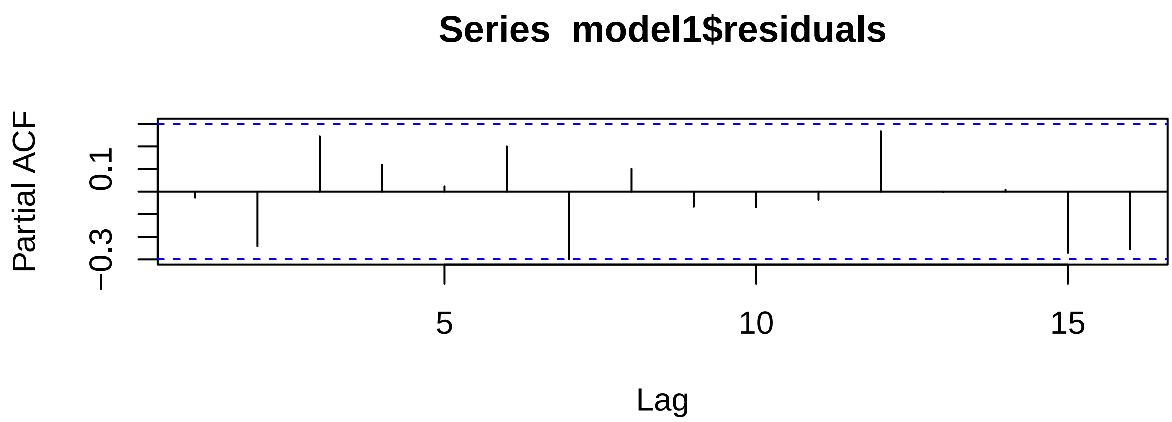
The sample EACF suggests that ARMA(2, 1) may be a good choice, so we can pick up this model first and try to estimate its parameters.

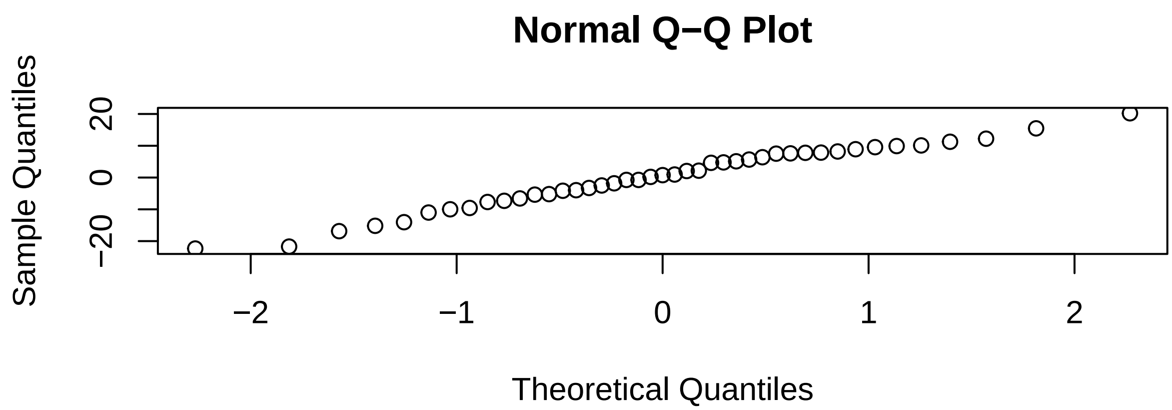
3.1.3 Estimation

```
##
## Call:
## stats::arima(x = O3, order = c(2, 0, 1), include.mean = T)
##
## Coefficients:
##          ar1      ar2      ma1  intercept
##       1.6803  -0.9291  -0.8189    57.6795
## s.e.  0.0458   0.0427   0.1019     1.2475
##
## sigma^2 estimated as 93.37:  log likelihood = -160.41,  aic = 330.81
```

3.1.4 Diagnostic

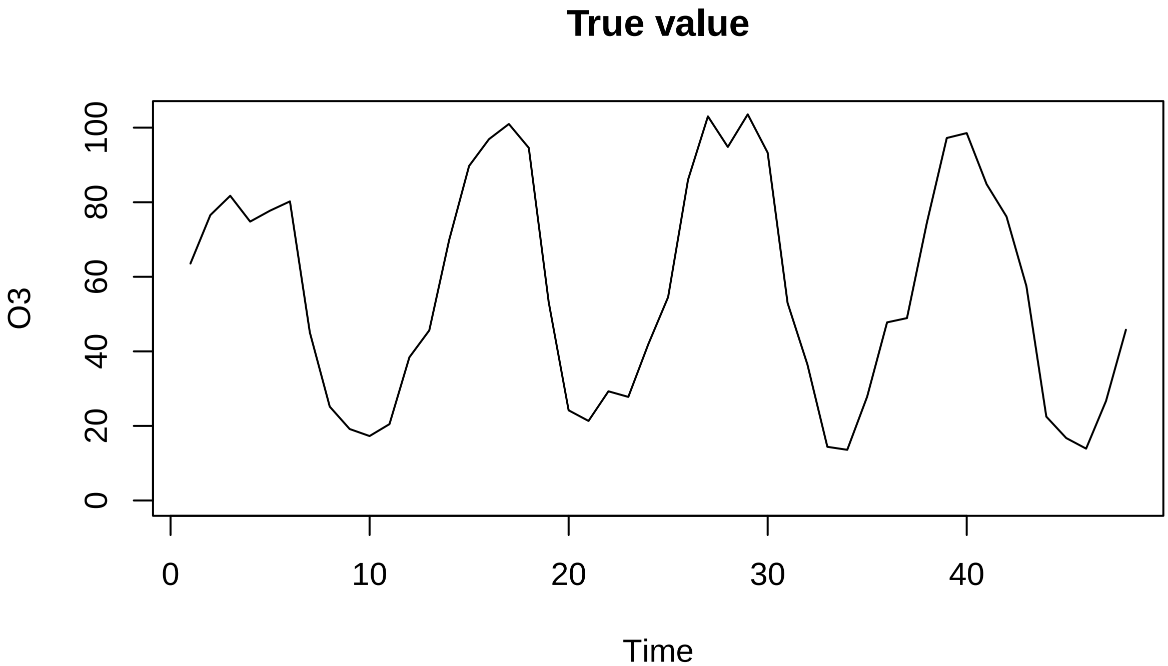


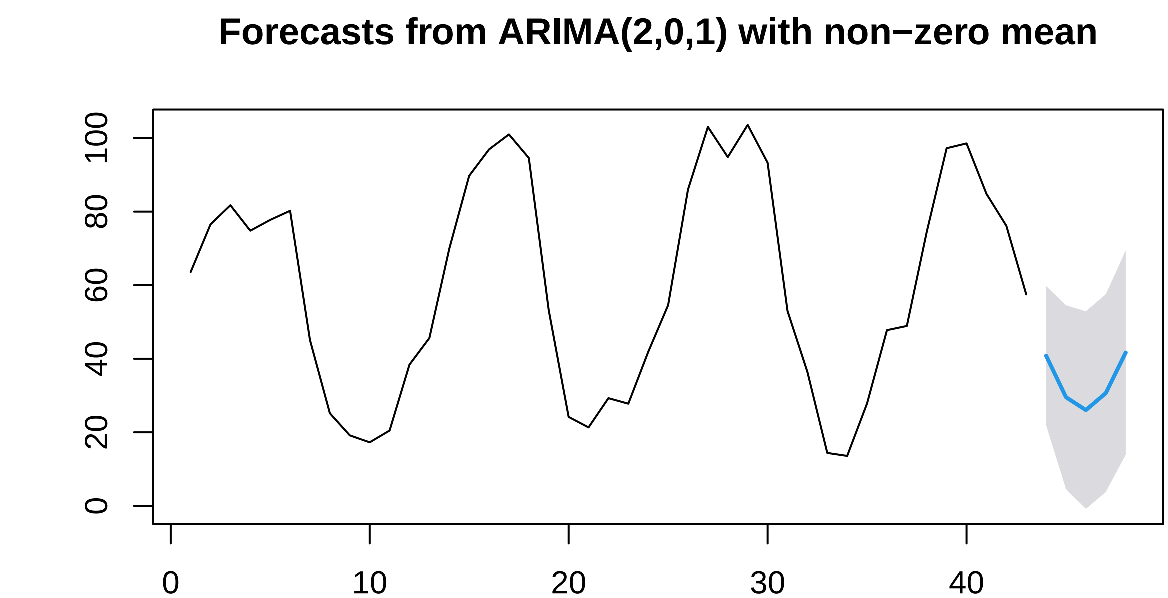




From the normal Q-Q plot, we can see that the residuals almost satisfy normal distribution. But from the ACF and PACF of the residuals, we may find that at some lag, the value is a little bit large, about 0.3 or so, which is not a really big problem. Also, the p-value of the Box test at lag 3 and 4 are close to 0.05, which means the null hypothesis is not that reliable.

3.1.5 Forecasting and Validation





The true value have more drastic change than the predicted value, therefore a better model should be considered.

3.2 Seasonal ARIMA Model for `O3` Data

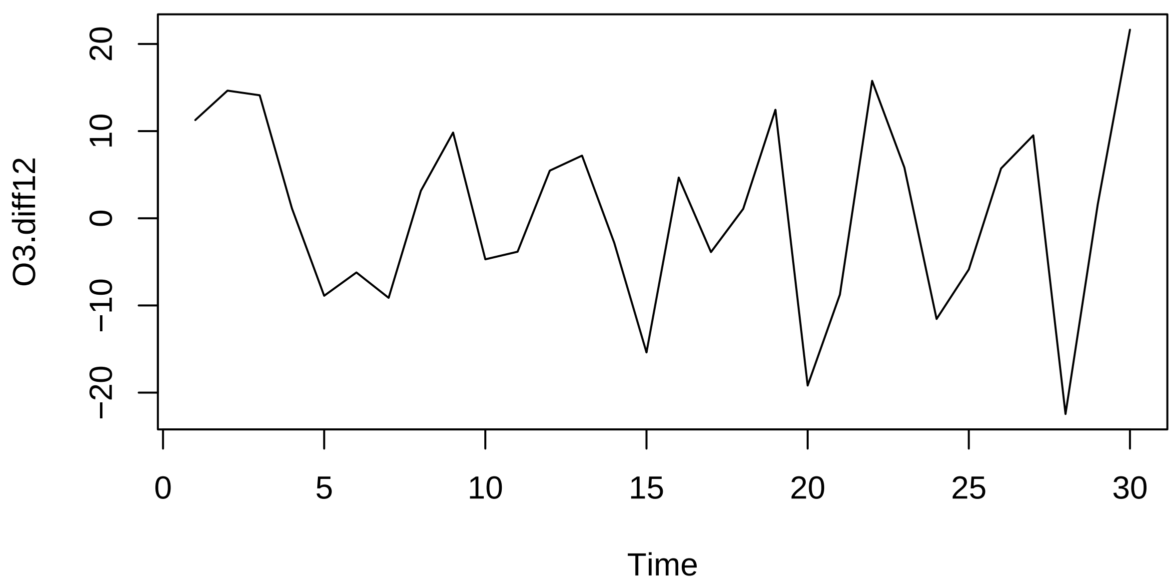
From empirical knowledge and the sequence plot above, seasonal ARIMA model seems to be a better choice to analyse this problem, since the data shows a seasonal period about 12.

3.2.1 Transformation and Stationarity Test

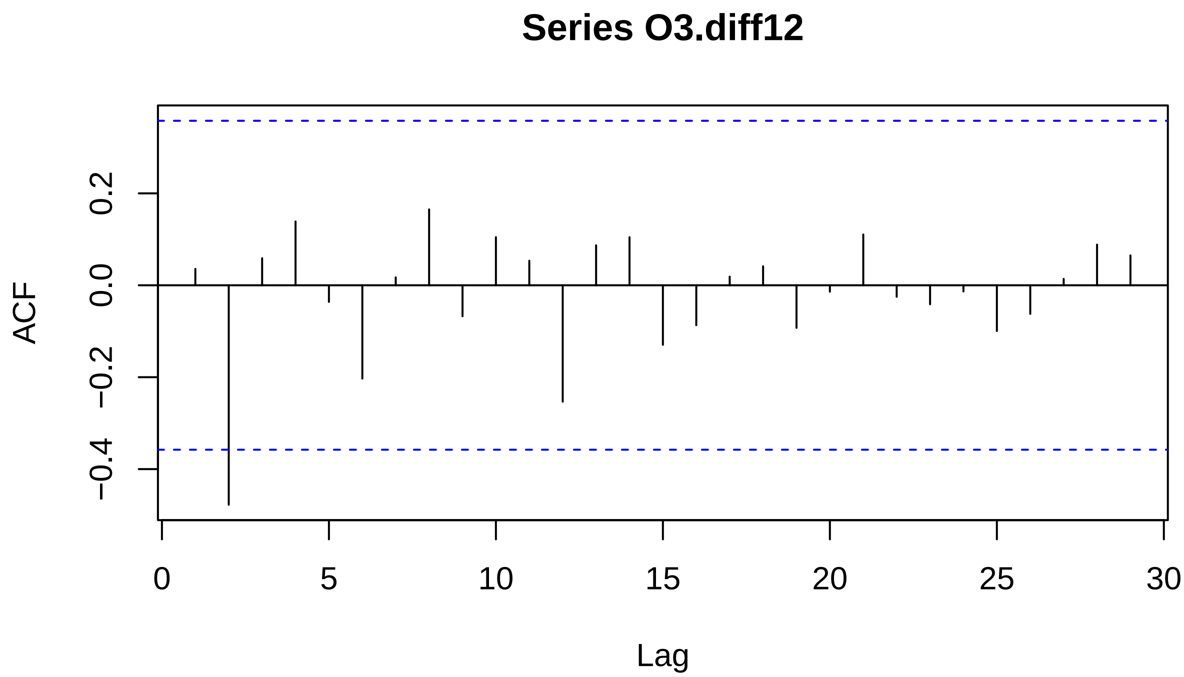
After differencing the sequence at lag 1 and lag 12 (i.e. $(1-B^{12})(1-B)y_t$), the new sequence is stationary.

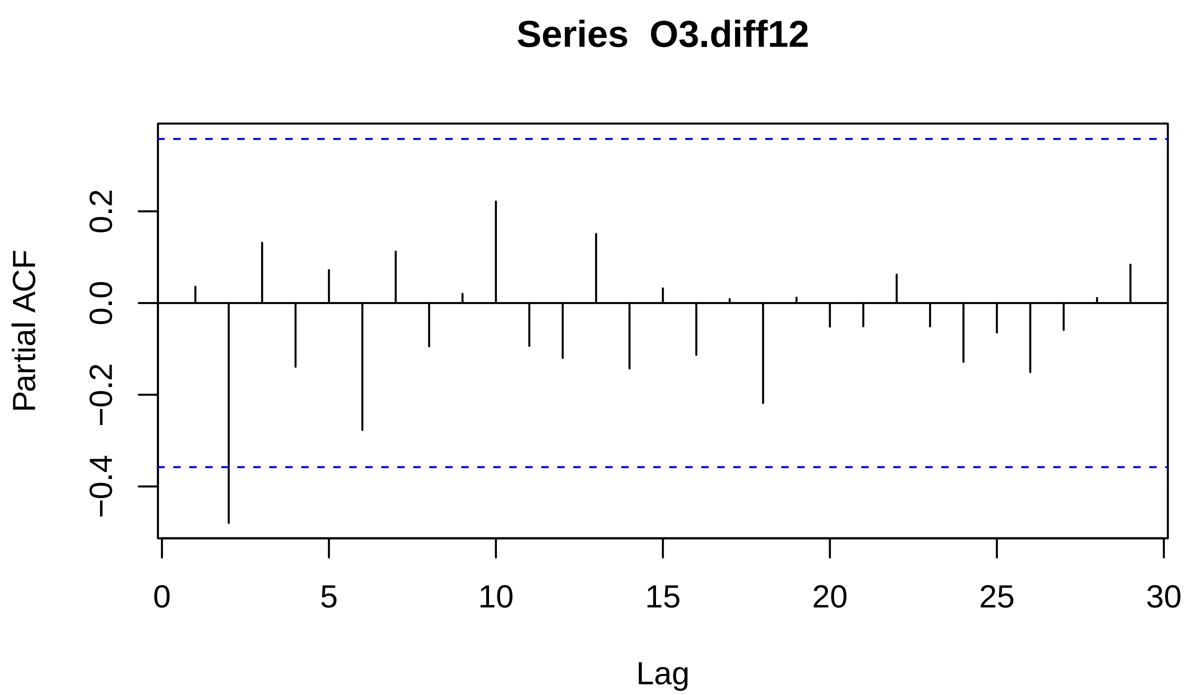
```
## Warning in adf.test(O3.diff12): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  O3.diff12
## Dickey-Fuller = -4.7148, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```



3.2.2 Specification





Here we have chosen $period = 12$, and $d = D = 1$. From the ACF plot, it has relatively large value at lag 1 and lag 12, therefore the model contains MA(1) component and seasonal MA(1) component. From the PACF plot, only at lag 1 do we find a relatively large value, therefore the model contains AR(1) component.

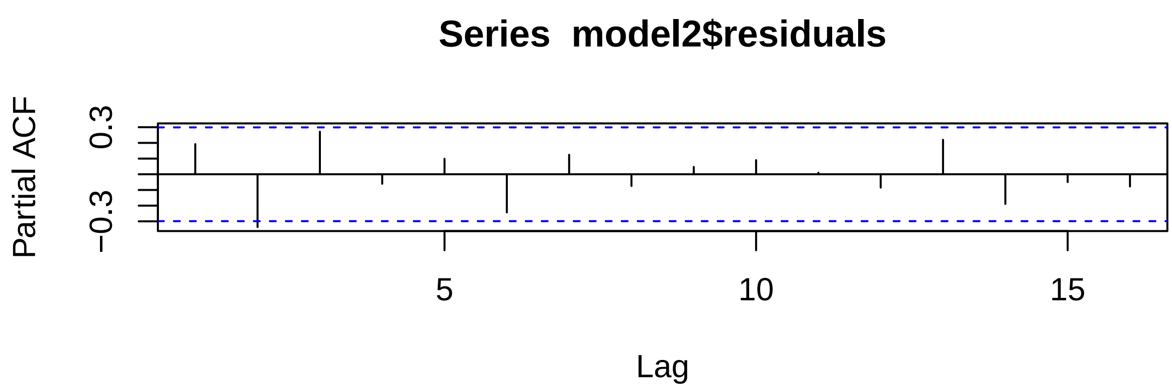
In conclusion, ARIMA(1, 1, 1)$\times$(0, 1, 1)$_{12}$ should be considered the possible proper model.

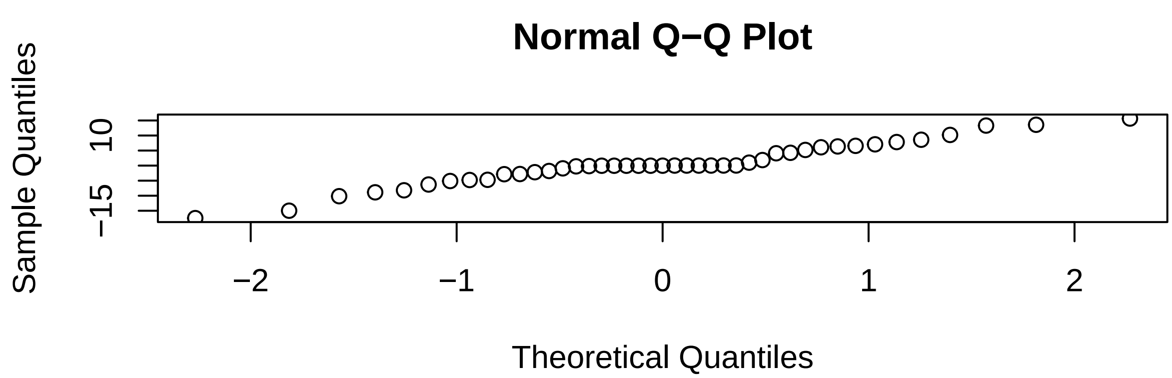
3.2.3 Estimation

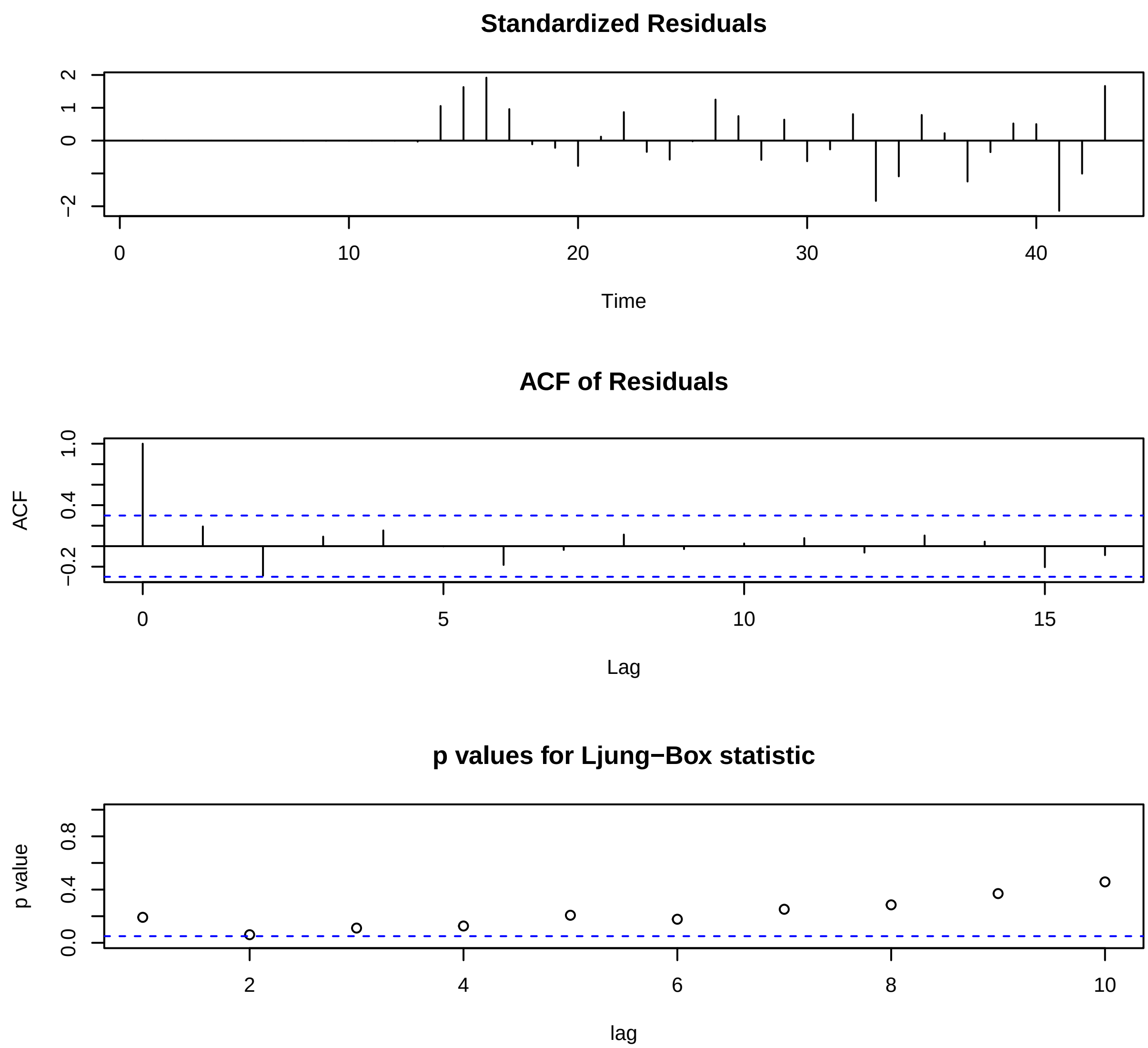
```
##
## Call:
## stats::arima(x = O3, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1),
##     period = 12), include.mean = T)
##
## Coefficients:
##          ar1      ma1     sma1
##       0.5747  -1.0000  -0.5922
## s.e.  0.1739   0.1914   0.4074
##
## sigma^2 estimated as 66.57:  log likelihood = -109.37,  aic = 226.75
```

Here, the AIC is much less than the ordinary ARIMA model in the last part, and should be a better model.

3.2.4 Diagnostic

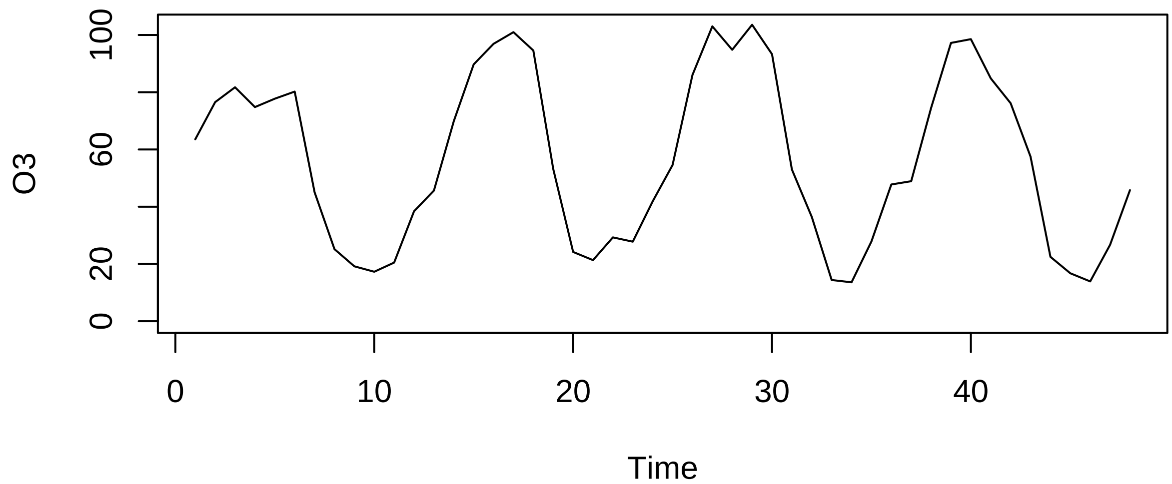


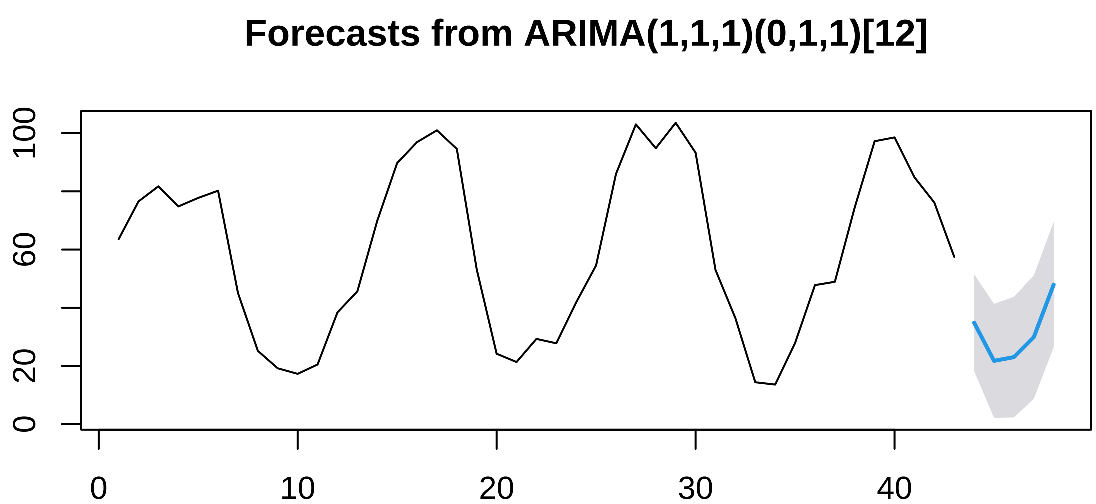




The ACF and PACF still have some relatively large value, and the normal Q-Q plot seems worse than the ARIMA model. At lag 2, the p-value of Box test is almost 0.05, which is not that good. But we should see the forecasting part to judge whether the model is effective.

3.2.5 Forecasting and Validation

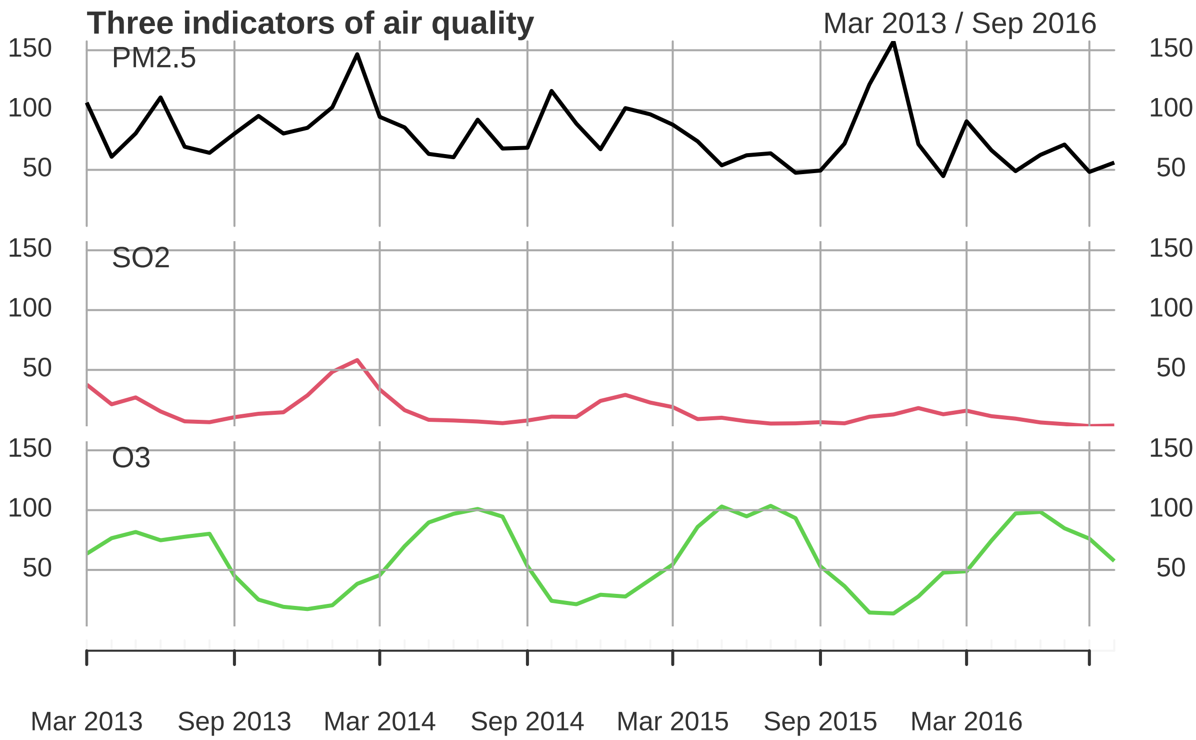




The forecasting gives us a promising result, showing drastic change like true value.

3.3 VAR Model for `PM2.5`, `SO2` and `O3` Data

Since there are so many indicators of air quality, we are also interested in whether they (or part of them) are correlated. Luckily, VAR model may be a powerful tool for us to get a deep insight into the correlation of the indicators. In this part, we use three pieces of data, which are `PM2.5`, `SO2` and `O3`, to analyse their potential correlation. The sequences are plotted as follows.



3.3.1 Specification

The information of VAR(1) to VAR(5) are listed as follows.

```
## Warning:   'MTS' R 4.1.3

## selected order: aic =  4
## selected order: bic =  2
## selected order: hq =  2
## Summary table:
##       p     AIC     BIC      HQ    M(p) p-value
## [1,] 0 17.3175 17.3175 17.3175  0.0000  0.0000
## [2,] 1 15.2629 15.6315 15.3988 82.8518  0.0000
## [3,] 2 14.4501 15.1874 14.7220 37.5571  0.0000
## [4,] 3 14.5372 15.6431 14.9450  9.1164  0.4266
## [5,] 4 14.3033 15.7778 14.8470 15.9876  0.0671
## [6,] 5 14.3871 16.2302 15.0667  7.1988  0.6164
```

Here, though the result of AIC suggests that VAR(4) may be good, but the p-value tells us that it may not be that significant. Therefore, VAR(2) is selected.

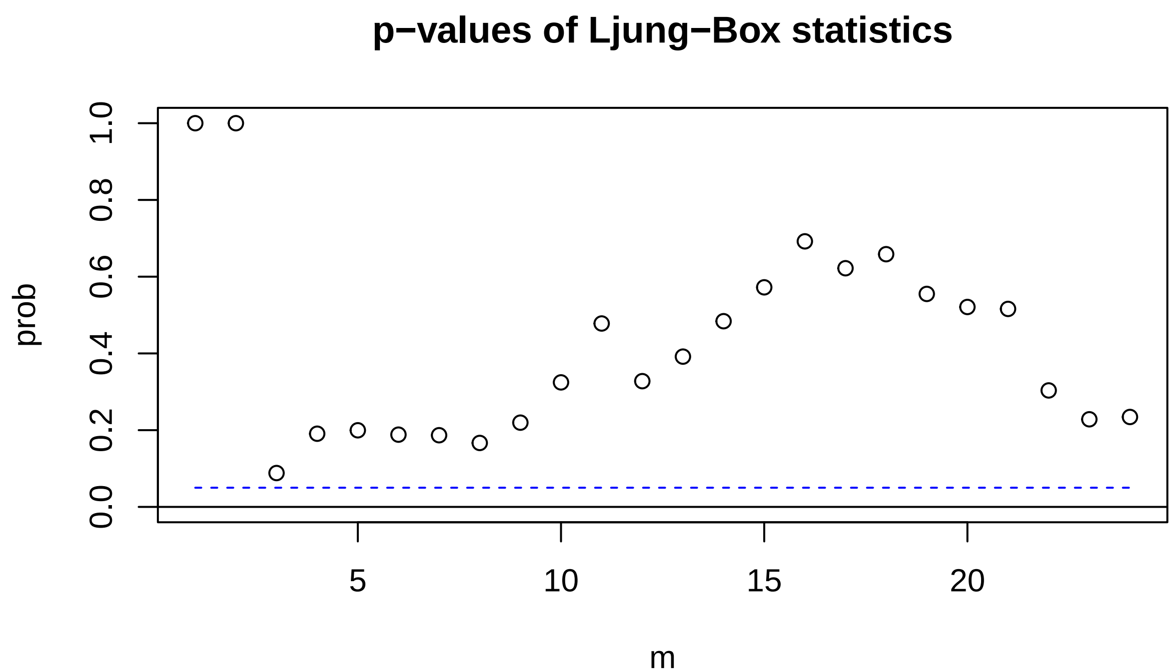
3.3.2 Estimation

```
## Constant term:
## Estimates:  122.5135 22.52475 11.11147
## Std.Error:  24.91042 8.063954 16.06137
## AR coefficient matrix
## AR( 1 )-matrix
##         [,1]   [,2]    [,3]
## [1,] -0.0302  0.901 -0.6524
## [2,] -0.1043  0.953 -0.0961
## [3,]  0.1365 -0.503  1.3516
## standard error
##        [,1]  [,2]   [,3]
## [1,] 0.1565 0.503 0.2109
## [2,] 0.0507 0.163 0.0683
## [3,] 0.1009 0.325 0.1360
## AR( 2 )-matrix
##         [,1]    [,2]    [,3]
## [1,] -0.3988  0.0903  0.2737
## [2,]  0.0105 -0.3447 -0.0611
## [3,] -0.0885  0.6336 -0.6507
## standard error
##        [,1]  [,2]   [,3]
## [1,] 0.1683 0.476 0.2323
## [2,] 0.0545 0.154 0.0752
## [3,] 0.1085 0.307 0.1498
##
## Residuals cov-mtx:
##           [,1]      [,2]      [,3]
## [1,] 287.32781  22.52830 -43.53453
## [2,]  22.52830  30.11007 -10.95958
## [3,] -43.53453 -10.95958 119.44846
##
## det(SSE) =  902700
## AIC =  14.55035
## BIC =  15.2876
## HQ  =  14.82223
```

3.3.3 Diagnostic

```
## Ljung-Box Statistics:
##          m       Q(m)     df    p-value
##  [1,]   1.00      4.84   -9.00     1.00
##  [2,]   2.00      9.35    0.00     1.00
##  [3,]   3.00     15.10    9.00     0.09
##  [4,]   4.00     23.00   18.00     0.19
##  [5,]   5.00     32.92   27.00     0.20
##  [6,]   6.00     43.28   36.00     0.19
##  [7,]   7.00     53.24   45.00     0.19
##  [8,]   8.00     63.95   54.00     0.17
##  [9,]   9.00     71.37   63.00     0.22
## [10,]  10.00     76.91   72.00     0.32
## [11,]  11.00     81.04   81.00     0.48
```

```
## [12,]  12.00    95.43   90.00     0.33
## [13,]  13.00   102.24   99.00     0.39
## [14,]  14.00   107.92  108.00     0.48
## [15,]  15.00   113.58  117.00     0.57
## [16,]  16.00   117.56  126.00     0.69
## [17,]  17.00   129.31  135.00     0.62
## [18,]  18.00   136.53  144.00     0.66
## [19,]  19.00   149.93  153.00     0.56
## [20,]  20.00   160.39  162.00     0.52
## [21,]  21.00   169.60  171.00     0.52
## [22,]  22.00   189.25  180.00     0.30
## [23,]  23.00   203.15  189.00     0.23
## [24,]  24.00   212.08  198.00     0.23
```



The Box test shows that it's safe to say that the residuals are white noise, and the model can be used confidently.

3.3.4 Simplification

As the VAR(2) needs too many parameters, we may restrict some non-significant parameters to 0 to decrease the complexity of the model. Here, we set the threshold as 1.96 (5% significance level for t-test).

```
## Constant term:
## Estimates:  133.84 22.20171 0
## Std.Error:  14.35784 5.570016 0
## AR coefficient matrix
## AR( 1 )-matrix
##        [,1]  [,2]   [,3]
```

```
## [1,]  0.000 0.789 -0.459
## [2,] -0.109 0.961 -0.147
## [3,]  0.156 0.000  1.457
## standard error
##        [,1]  [,2]   [,3]
## [1,] 0.0000 0.288 0.1110
## [2,] 0.0496 0.149 0.0423
## [3,] 0.0365 0.000 0.1101
## AR( 2 )-matrix
##        [,1]  [,2]   [,3]
## [1,] -0.472  0.00  0.000
## [2,]  0.000 -0.29  0.000
## [3,]  0.000  0.00 -0.675
## standard error
##       [,1]  [,2]  [,3]
## [1,] 0.127 0.000 0.000
## [2,] 0.000 0.128 0.000
## [3,] 0.000 0.000 0.112
##
## Residuals cov-mtx:
##           [,1]      [,2]      [,3]
## [1,] 299.20166  19.80839 -41.15502
## [2,]  19.80839  30.99733 -12.18621
## [3,] -41.15502 -12.18621 134.54976
##
## det(SSE) =  1118017
## AIC =  14.39218
## BIC =  14.80176
## HQ  =  14.54322
```

The AIC and BIC of the simplified model are both smaller than the primal model, therefore the new one is better. Note that here 9 parameters have been restricted to 0.

3.3.5 Granger Causality Test

We can also do Granger causality test based on the model above.

```
## [1] "PM2.5"

## Number of targeted zero parameters:  4
## Chi-square test for Granger Causality and p-value:  25.2801 4.418878e-05

## [1] "SO2"

## Number of targeted zero parameters:  4
## Chi-square test for Granger Causality and p-value:  13.16048 0.0105177

## [1] "O3"

## Number of targeted zero parameters:  4
## Chi-square test for Granger Causality and p-value:  7.084817 0.1314734
```

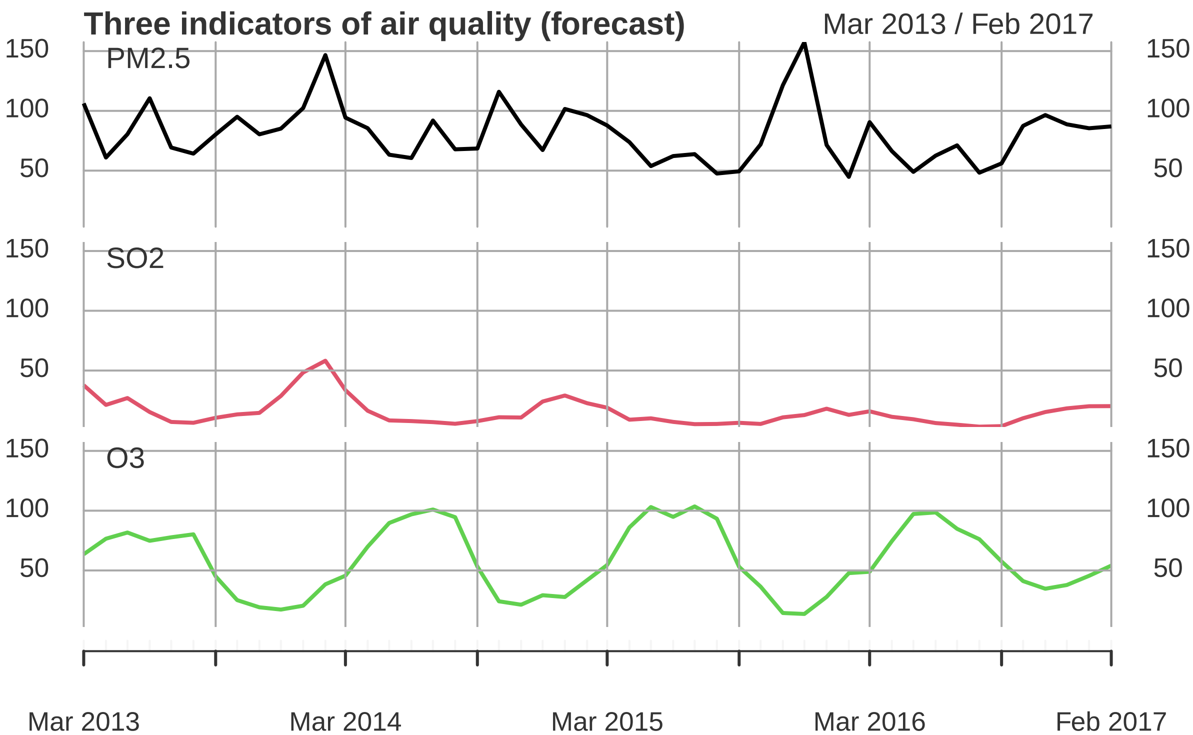
```
## Constant term:
## Estimates:  66.66082 22.52475 11.11147
## Std.Error:  14.93489 8.063954 16.06137
## AR coefficient matrix
## AR( 1 )-matrix
##        [,1]   [,2]    [,3]
## [1,]  0.452  0.000  0.0000
## [2,] -0.104  0.953 -0.0961
## [3,]  0.137 -0.503  1.3516
## standard error
##        [,1]  [,2]   [,3]
## [1,] 0.1534 0.000 0.0000
## [2,] 0.0507 0.163 0.0683
## [3,] 0.1009 0.325 0.1360
## AR( 2 )-matrix
##         [,1]   [,2]    [,3]
## [1,] -0.2846  0.000  0.0000
## [2,]  0.0105 -0.345 -0.0611
## [3,] -0.0885  0.634 -0.6507
## standard error
##        [,1]  [,2]   [,3]
## [1,] 0.1545 0.000 0.0000
## [2,] 0.0545 0.154 0.0752
## [3,] 0.1085 0.307 0.1498
##
## Residuals cov-mtx:
##           [,1]      [,2]      [,3]
## [1,] 500.96533  22.52830 -43.53453
## [2,]  22.52830  30.11007 -10.95958
## [3,] -43.53453 -10.95958 119.44846
##
## det(SSE) =  1645408
## AIC =  14.96466
## BIC =  15.53808
## HQ  =  15.17612
```

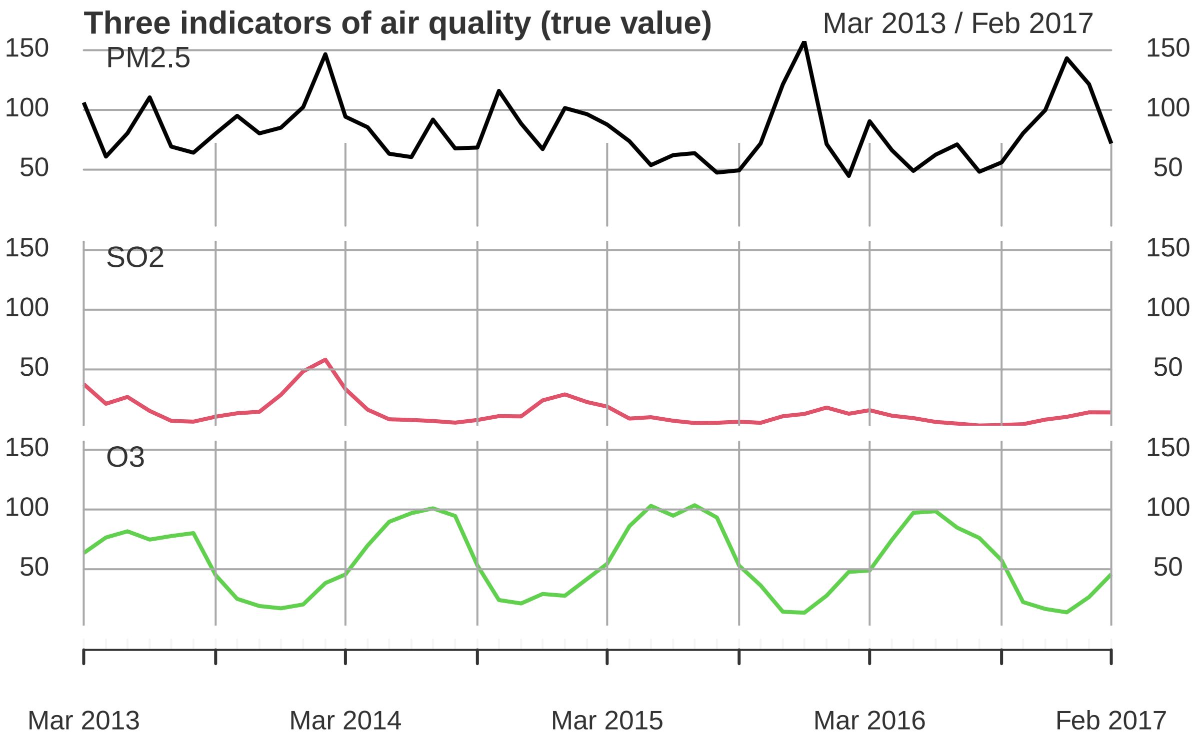
From the results above, the conclusions should be: at 5% significance level, `O3` can be regarded as the one-way Granger cause of `SO2` and `PM2.5`.

3.3.6 Forecasting and Validation

```
## orig  43
## Forecasts at origin:  43
##      PM2.5   SO2    O3
## [1,] 87.43 10.11 41.10
## [2,] 96.49 15.31 34.69
## [3,] 88.75 18.35 37.82
## [4,] 85.44 20.15 45.50
## [5,] 86.98 20.23 54.05
## Standard Errors of predictions:
##       [,1]   [,2]  [,3]
## [1,] 17.30  5.568 11.60
## [2,] 18.86  8.003 20.21
```

```
## [3,] 23.02  9.426 25.50
## [4,] 24.27 10.388 27.43
## [5,] 24.88 11.022 27.64
## Root mean square errors of predictions:
##          [,1]     [,2]    [,3]
## [1,]   18.65    6.004   12.51
## [2,] 3054.10 2336.011 6726.78
## [3,] 5361.29 2023.499 6318.30
## [4,] 3132.50 1774.035 4102.17
## [5,] 2223.53 1497.800 1403.25
```





Only in the sequence of `PM2.5` do we have a relatively big bias, the other 2 sequences are predicted well. But it's understandable since the true value of `PM2.5` sequence in our prediction window is a "peak" and is really hard to predict.

4 Conclusion

This project built a ARIMA model and a seasonal ARIMA model for the `O3` data, and found that the seasonal ARIMA model performed better at forecasting. Also, through VAR model, we found that `O3` can be regarded as the one-way Granger cause of `SO2` and `PM2.5`.

5 Reference

1. Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. and Chen, S.X. (2017) Cautionary Tales on Air-Quality Improvement in Beijing. Proceedings of the Royal Society A, Volume 473, No. 2205, Pages 20170457.

2. Dongfeng Li, Lecture Notes of Financial Time Series Analysis, can be accessed with: *https://www.math.pku.edu.cn/teachers/lidf/course/fts/ftsnotes/html/_ftsnotes/index.html*.