Beijing Multi-Site Air-Quality Data Set

Danyu Zhang & Daniel Alonso

March 21st, 2021

Description of Beijing Multi-Site Air-Quality Dataset

This 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. The meteorological data in each air-quality site is matched with the nearest weather station from the China Meteorological Administration. The time period starts on March 1st, 2013 and ends on February 28th, 2017 (35064 observations). Missing data are denoted as NA. For more information please check this link.

For this project, we will only consider district Tiantan (Temple of Heaven) of Beijing, which is a very centric and popularly visited zone in the city and the district of Dingling, which is a suburban district. Our purpose is to check if there's a difference in O_3 pollution throughout time for both districts.

The data set contains the the following variables:

- **year, month, day and hour**: Time variable which denotes the year, month, day and hour of the taken value for each meteorological variables the data set contains.
- **PM2.5**: Fine Suspended Particles, PM2.5 concentration (ug/mˆ3),
- **PM10**: Respirable suspended particulates, PM10 concentration (ug/mˆ3)
- **O3**: O_3 concentration (ug/mˆ3). Ozone is a gas composed of three atoms of oxygen (O_3).
- **SO2**: SO2 concentration (ug/mˆ3). Sulphur dioxide (SO2) is an air pollutant made up of sulphur and oxygen atoms and is harmful to both plants and people. On dissolution in rain water, SO_2 produces acid rain. This SO_3 gets converted into H2SO4 in the presence of moisture, which comes down in the form acid rain.
- **NO2**: NO2 concentration (ug/mˆ3). NO2 primarily gets in the air from the burning of fuel. NO_2 forms from emissions from cars, trucks and buses, power plants, and off-road equipment.
- **CO**: CO concentration (ug/mˆ3). Carbon Monoxide, the greatest sources of CO to outdoor air are cars, trucks and other vehicles or machinery that burn fossil fuels.

We will only be using the date variables and O_3 pollution level.

Simple view of the modified data set:

Table 1: Beijing Air-Quality Data Set

No	year	month	day	hour	O3_Tiantan	O3_Dingling
1	2013	3	1	0	81	82
2	2013	3	1	1	80	80
3	2013	3	1	2	75	79
4	2013	3	1	3	74	79
5	2013	3	1	4	70	81
6	2013	3	1	5	70	79

1. Dealing with the Missing Values

We check if there are missing values in different time series and how many there are.

```
## [1] 843
```

```
## [1] 1214
```

During this step, we will deal with the missing values by using interpolation tecniques. As a result, we will be replacing them by averaging the values of the two nearest rows and imputing the values with this result.

We check that there are indeed no more missing values.

```
## [1] 0
```

```
## [1] 0
```

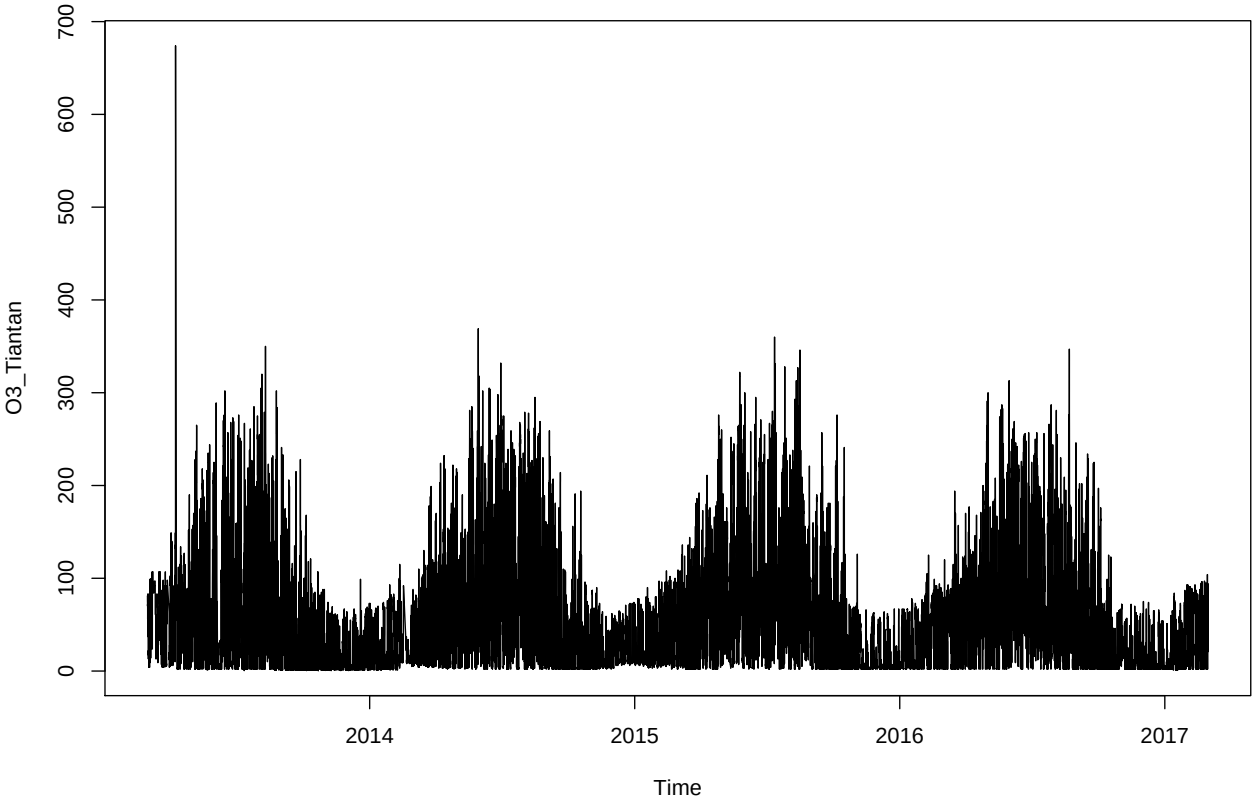
2. Visualizing time series

2.1 Original Hourly Data Set

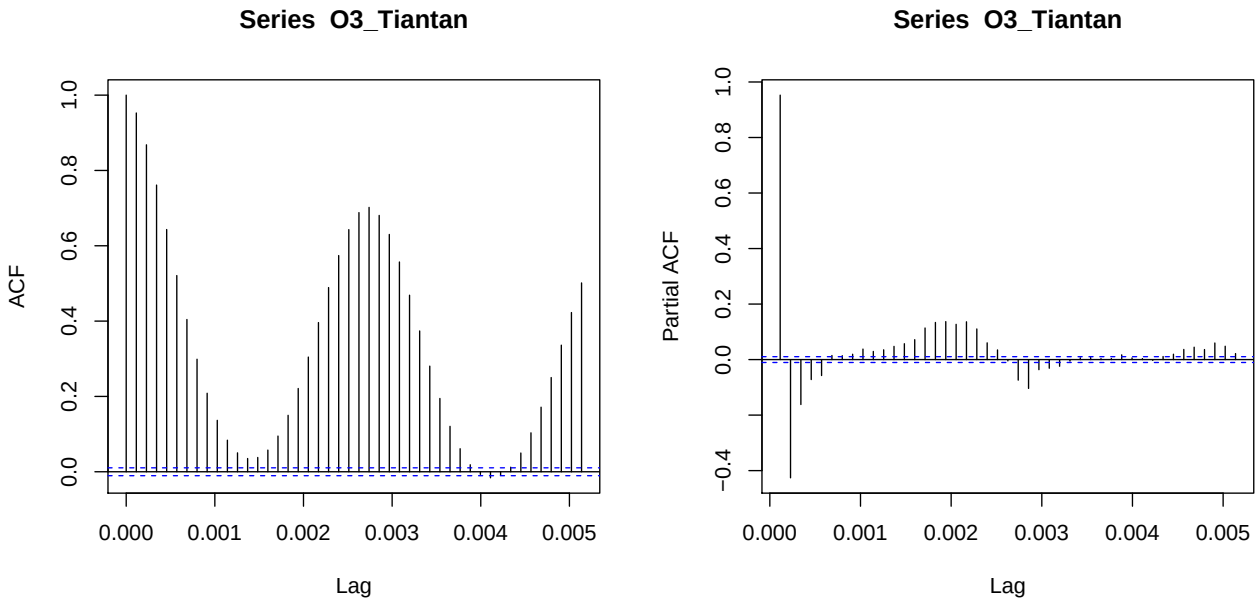
- **Tiantan District**

Apparently, in the case of the O_3 pollution time series in Tiantan, the TS is a non-stationary process: it does not have a constant mean or variance (heterocedastic process), also the covariance between the observations seems to not only depend on the lag, therefore the series has long-term memory.

Moreover, we can observe the seasonal variation pattern that repeats every year, during the middle of a year (summer) from 2013 to 2016, the ozone pollution increases significantly, while during the start of the year (late winter/early spring) the ozone pollution is lower, which could possibly be caused by the summer vacations, where people use cars more often. It is possible to eliminate that effect by subtracting the difference of 365 lags. Furthermore, there is a very significant outlier in the data set that seems to be an additive outlier due to the fact that it's only affecting 1 observation in the time series.



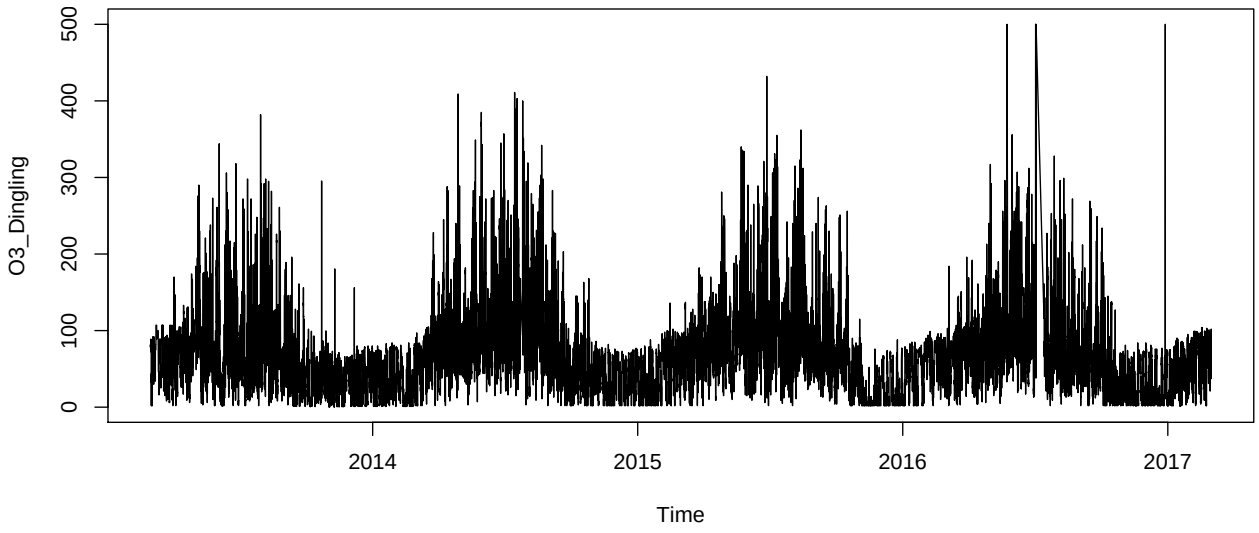
By checking the autocorrelation plot and the partial autocorrelation, it is very obvious that this process is not stationary, the autocorrelation decays very slowly and tops after another period of time, the partial autocorrelation has more or less the same pattern which indicates seasonalility.



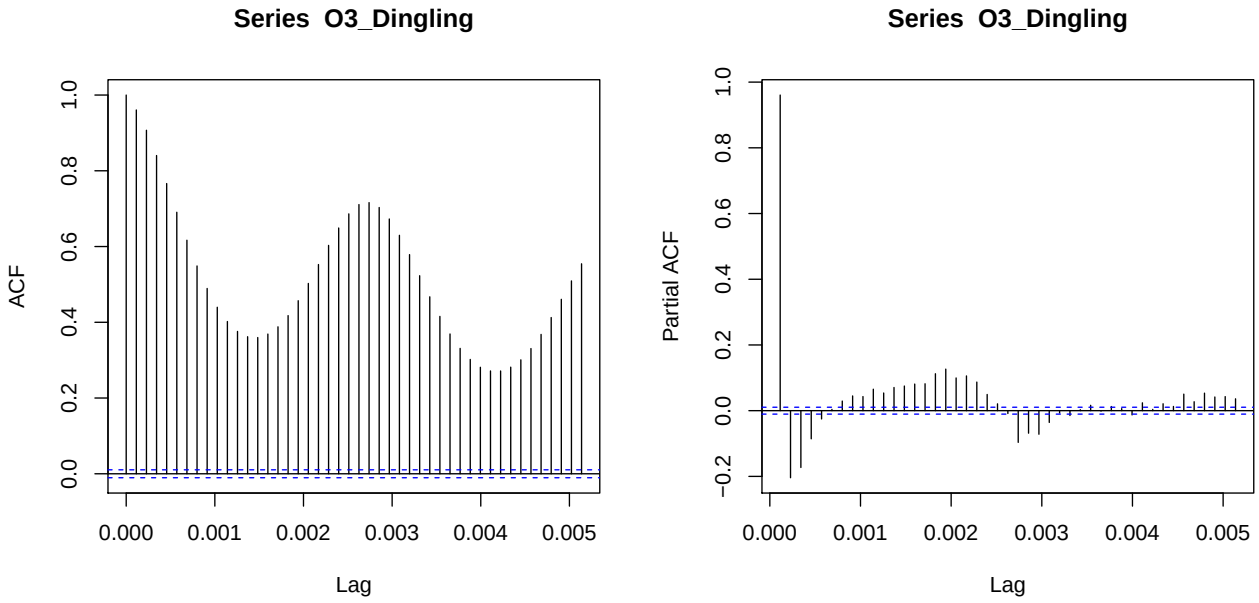
- **Dingling District**

The conclusions that we can tell from the plot are very similar to the plot before, it is not a stationary process due to the facts that it does not have a constant mean; nor a constant variance, in the Dingling district the series seems to vary more than in the Tiantan district. The autocorrelation does not only depend on the lag, as a result, it has long-term memory.

Also it shows a clear seasonal pattern, more ozone pollution during the summer and less ozone pollution during the winter. We will be able to eliminate that effect by calculating a difference of 365 lags. There are also significant outliers in the data set of Dingling District.



Very similar to the plots we obtained before for the Tiantan District, a non-stationary process (seasonal patterns).



2.2 Modified Daily Data Set

Given that we have hourly data, the amount of observations is massive. If we want to check more detailed patterns, it is going to be very laborious to do so. Although this could mean losing some data, we will average the 24 hourly observations in oder to obtain a daily mean ozone pollution time series.

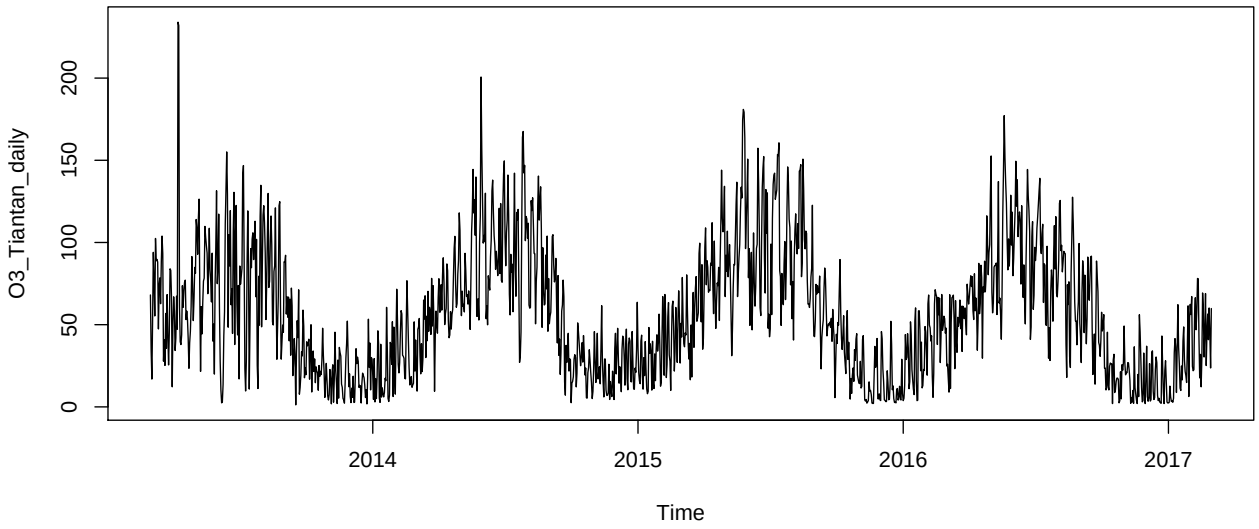
Table 2: Daily Beijing Air-Quality Data Set

date	mean_O3_Tiantan	mean_O3_Dingling
2013-3-1	68.08333	81.95833
2013-3-2	34.04167	24.70833
2013-3-3	16.91667	41.00000
2013-3-4	53.12500	76.91667
2013-3-5	94.00000	86.50000
2013-3-6	75.33333	73.12500

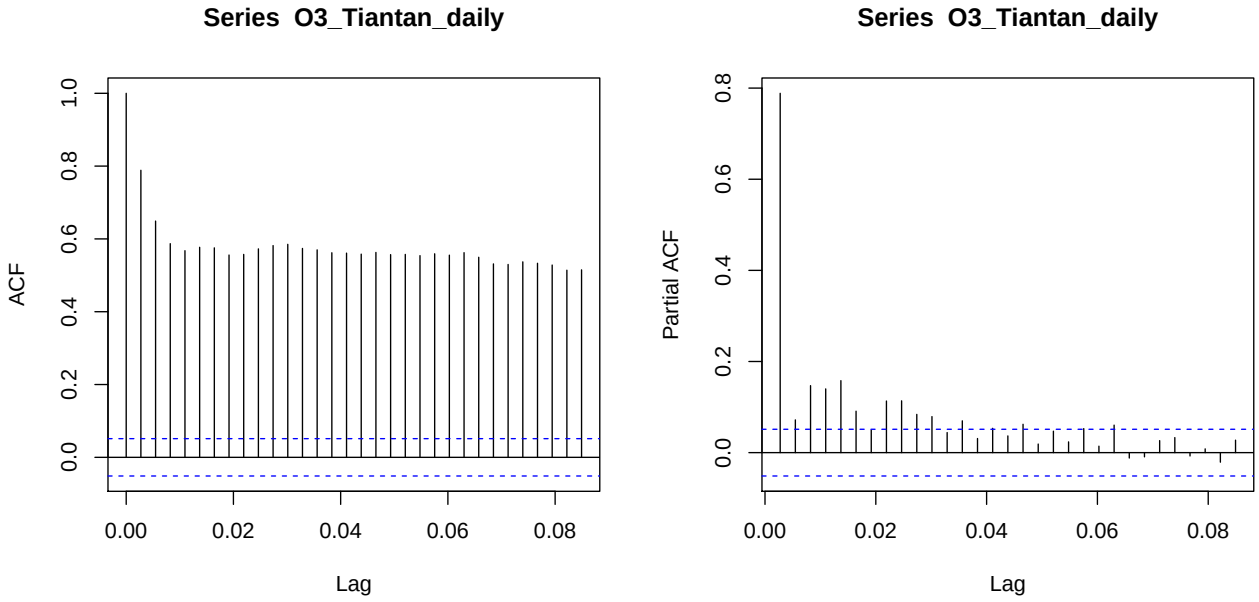
- **Tiantan District**

We can clearly observe that it's not stationary due to the fact that it does not have a constant mean, the variance seems to be more or less constant (homocedastic process), but the covariance between the observations seems to not only depend on the lag, as a result, this series also has long-term memory.

Additionaly, as we have concluded before in the hourly data set, there is seasonal variation that repeats every year, during the summer the ozone pollution is usually higher than during winter/spring. We are able to get rid of it by calculating 365 lag difference. Again, there is a very significant additive outlier, similar to the hourly data set (regardless of our transformation by converting the TS into a daily TS, averaging the values per hour seems to make no difference).



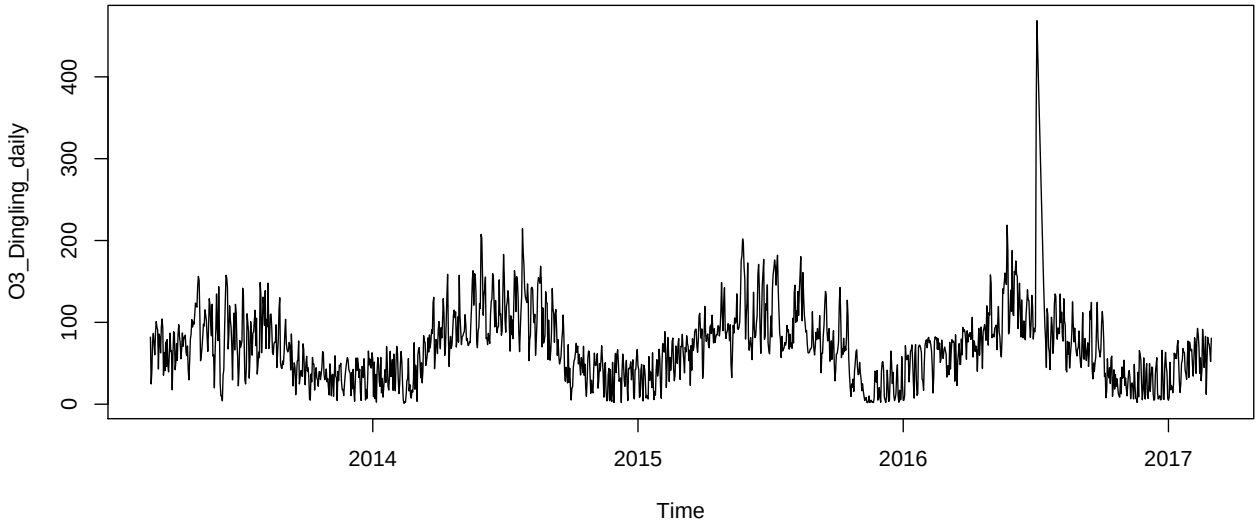
We can observe from the autocorrelation and the partial autocorrelation plots, that the series is not stationary, the autocorrelation decays very slowly and seems to stay at a point forever which indicates a seasonal variation, the partial autocorrelation has a lot of significant peaks too.



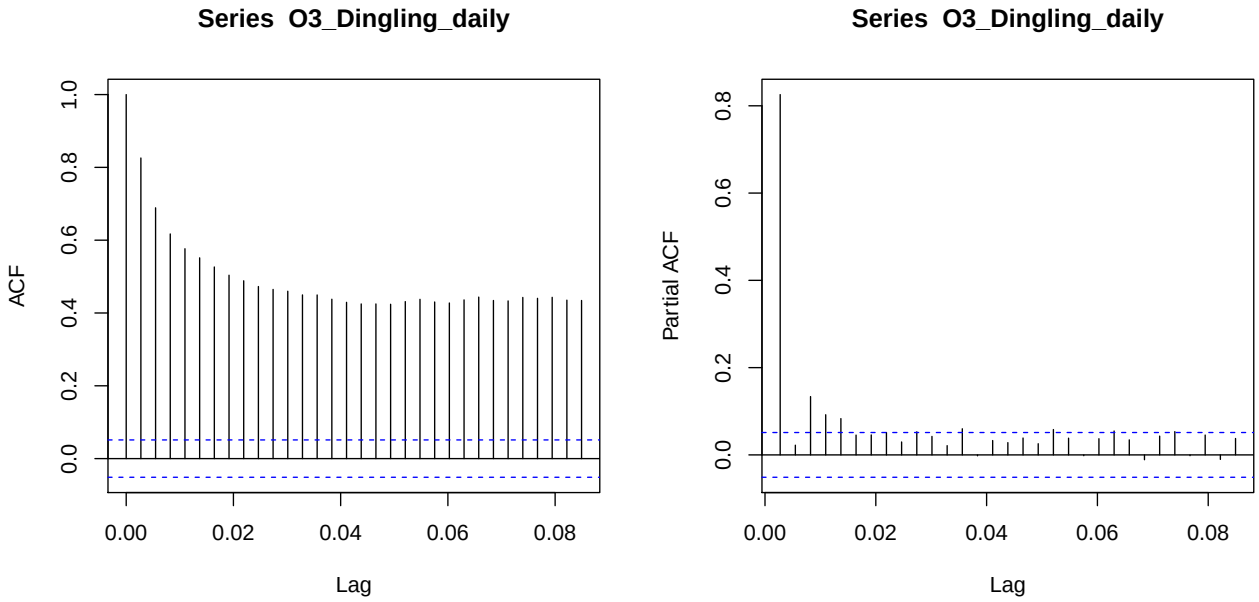
- **Dingling District**

From the following plot it is easy to notice that the process is again not stationary: the mean of the process not constant; the variance of the process seems to be constant; and the autocorrelation does not only depend on the lag, rendering it a long-term memory process.

Also it shows a clear seasonal pattern, more ozone pollution during middle of an year and less ozone pollution during start/end of an year. Elimination can be done by performing the 365 lag difference. There are also very significant outliers in the data set of the Dingling District that has even modified our y-axis scale.



This shows stark similarity to the plots we obtained before for Tiantan daily TS, a non-stationary process (seasonality patterns).



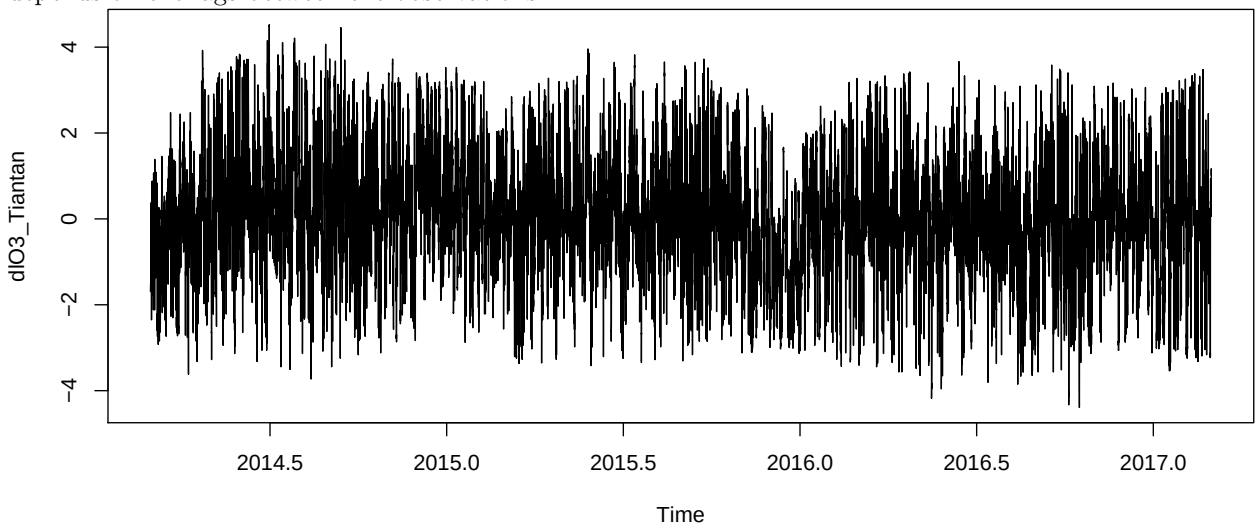
3. Obtaining Stationary Processes

3.1 Original Hourly Data Set

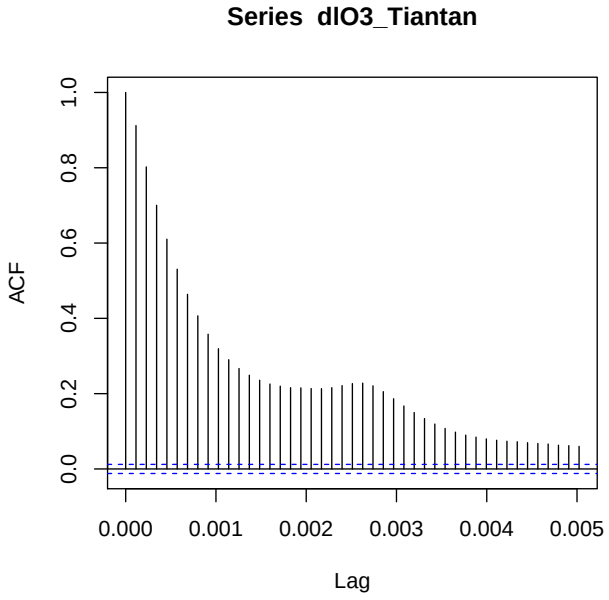
- **Tiantan District**

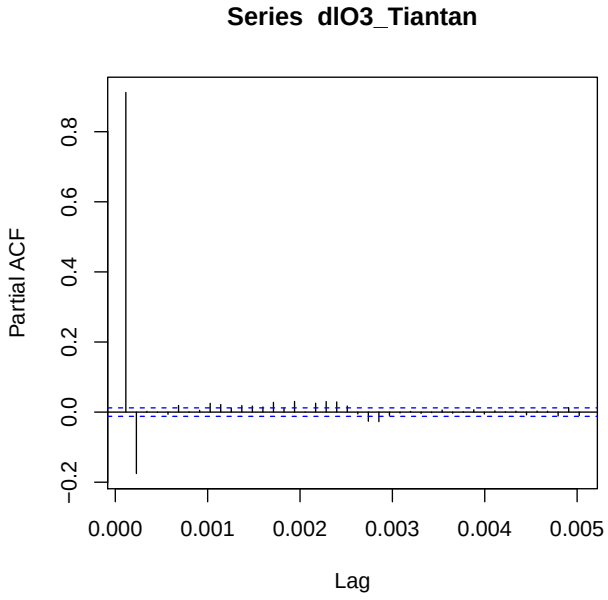
In order to obtain a stationary process for ozone pollution hourly data in Tiantan, first we eliminate the heterocedasticity by performing a log transformation, afterwards, we differenciate the process by 365*24 just to remove the yearly seasonal pattern.

We can observe that the process has constant mean aroud 0, constant variance, and the autocorrelation only depends on the lags between the observations.



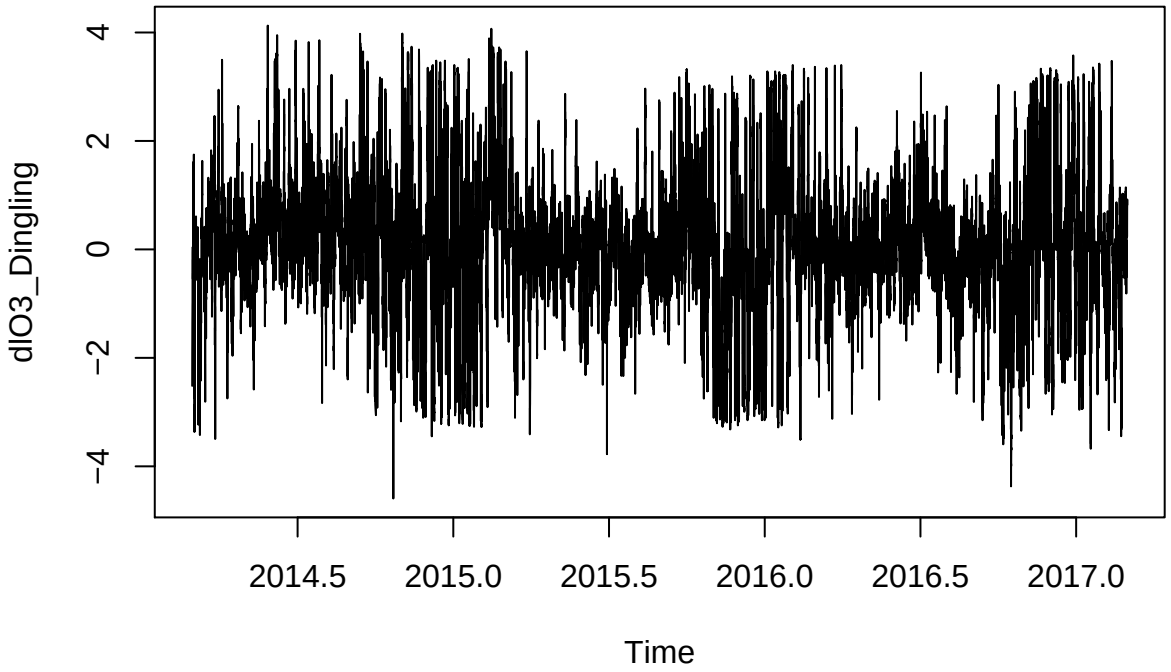
From the autocorrelation plot we can observe an exponentially decaying pattern, and from partial autocorrelation plot, we can only observe 2 siginificant peaks. This seems to be similar to a moving average process of order 2. We will later on fit the model in part 4.



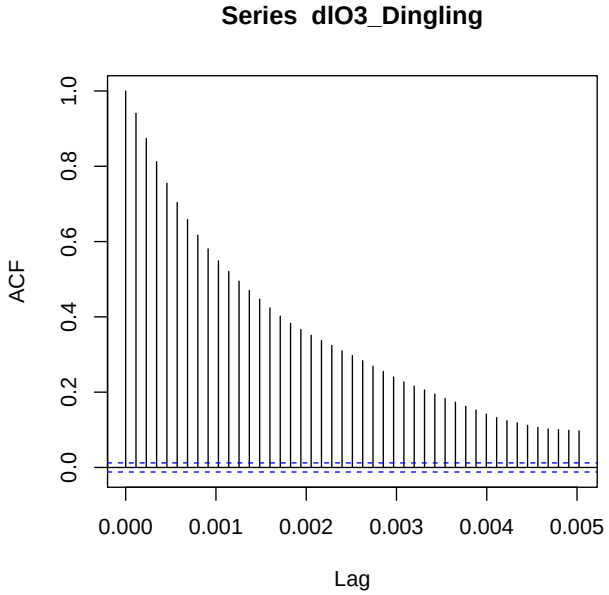


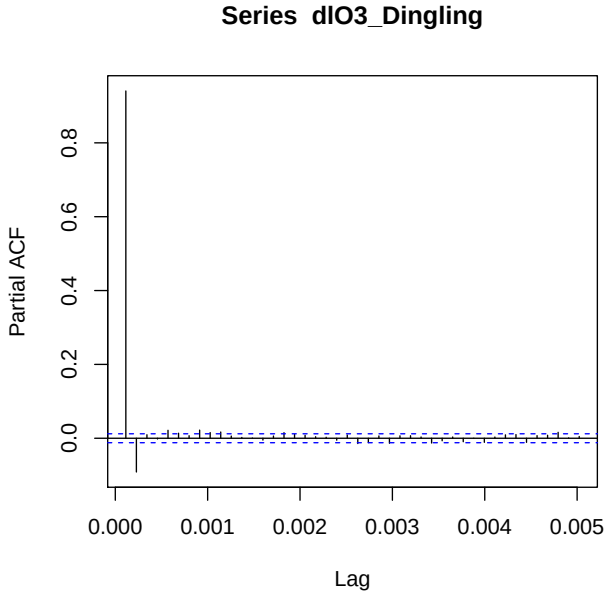
- **Dingling District**

For the Dingling District TS we do the exact same thing as we did for Tiantan District, after all the modifications are done, we can notice that it has constant mean, constant variance, and the autocorrelation only depends on the lags between the observations.



We have the same situation as before, the patterns are very similar to an MA(2) process.



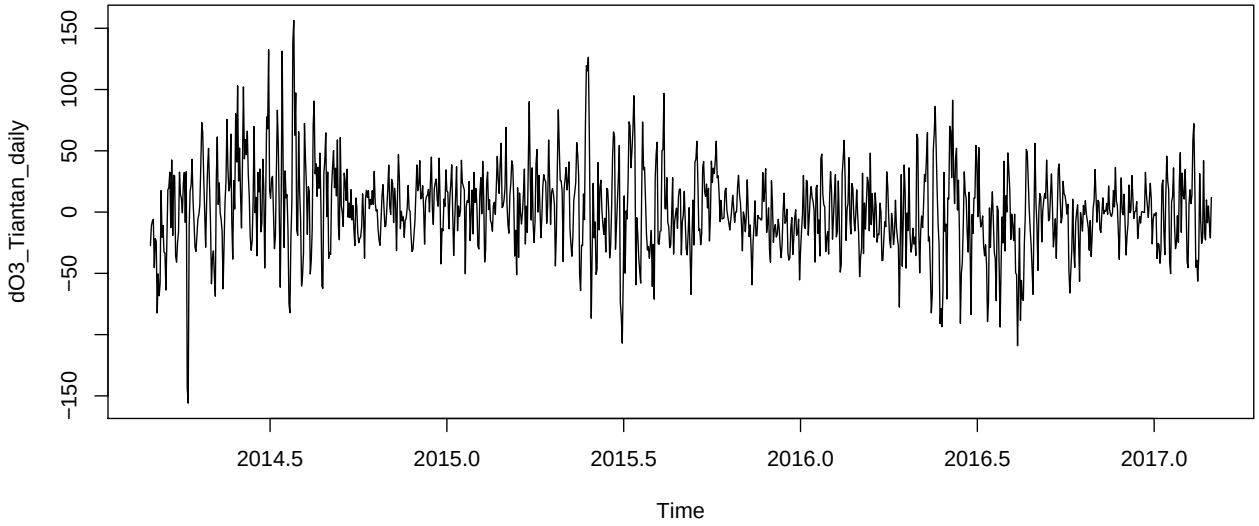


3.2 Modified Daily Data Set

- **Tiantan District**

In order to obtain a stationary process for the ozone pollution daily TS in Tiantan, it is essential to perform a 365 lag difference to remove the yearly seasonal pattern.

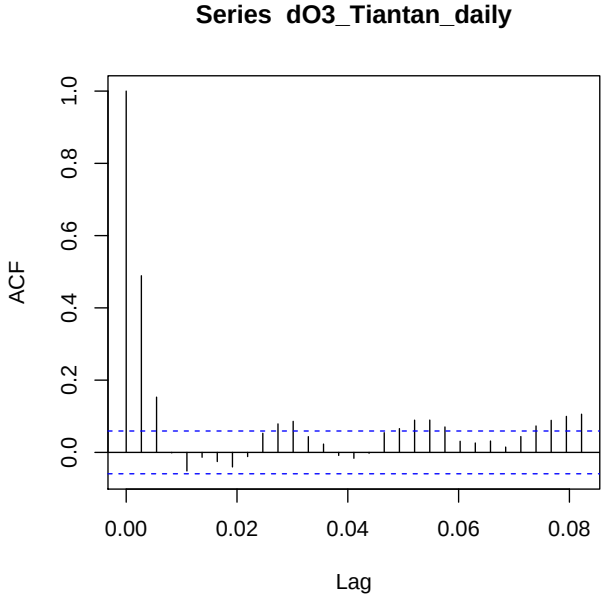
We can observe that the averaged ozone pollution data is more or less stationary: constant mean around 1, roughly constant variance with a variation, and the autocorrelation only depends on the lag between observations.

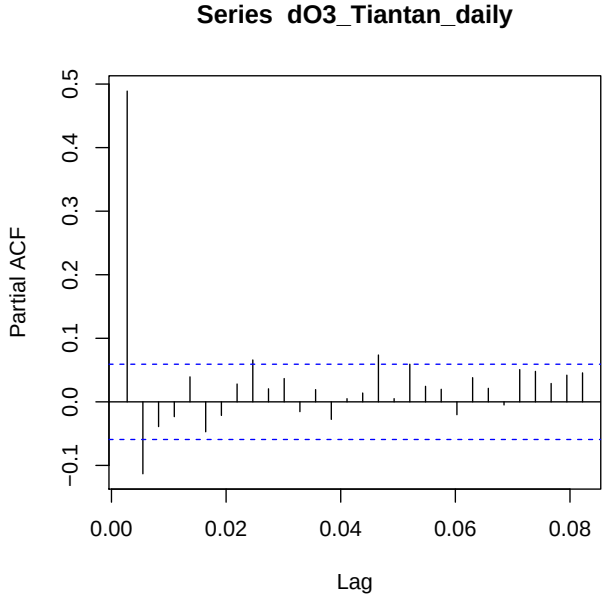


In the autocorrelation plot we can see that there are 2 significant peaks at first, and then some other significant peaks 11, 12, 20, 21... It is not possible to remove those significant peaks by applying a difference.

In the partial autocorrelation plot we can see that there are 2 significant peaks, and as in the ACF, some other significant peaks after that.

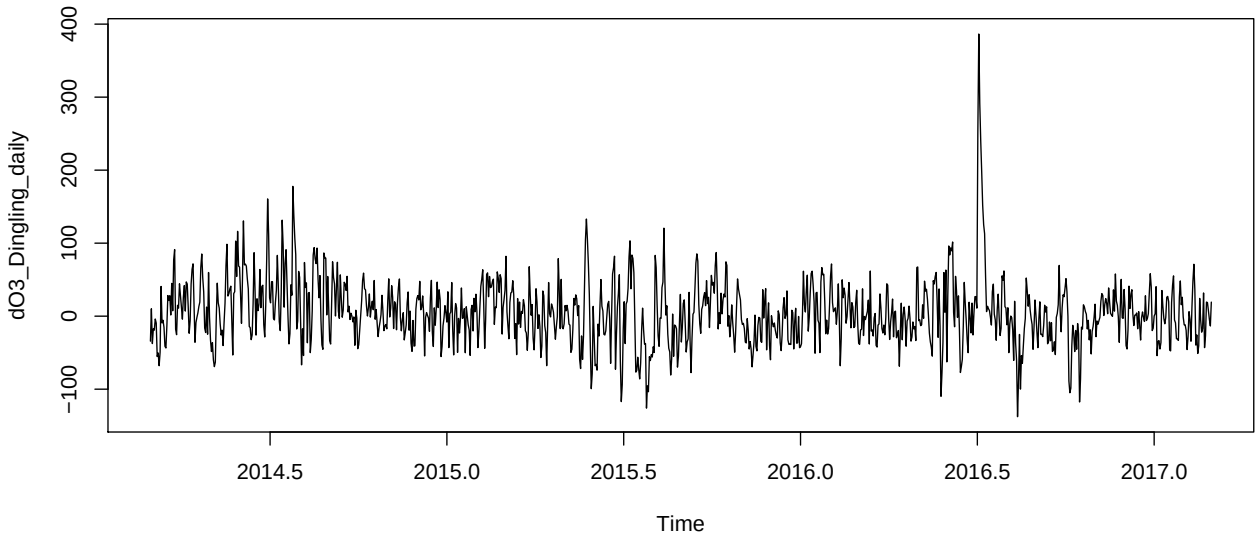
By observing the ACF and PACF, it is likely that the ozone daily pollution in Tiantan follows ARIMA(2,0,2) model in the first place.





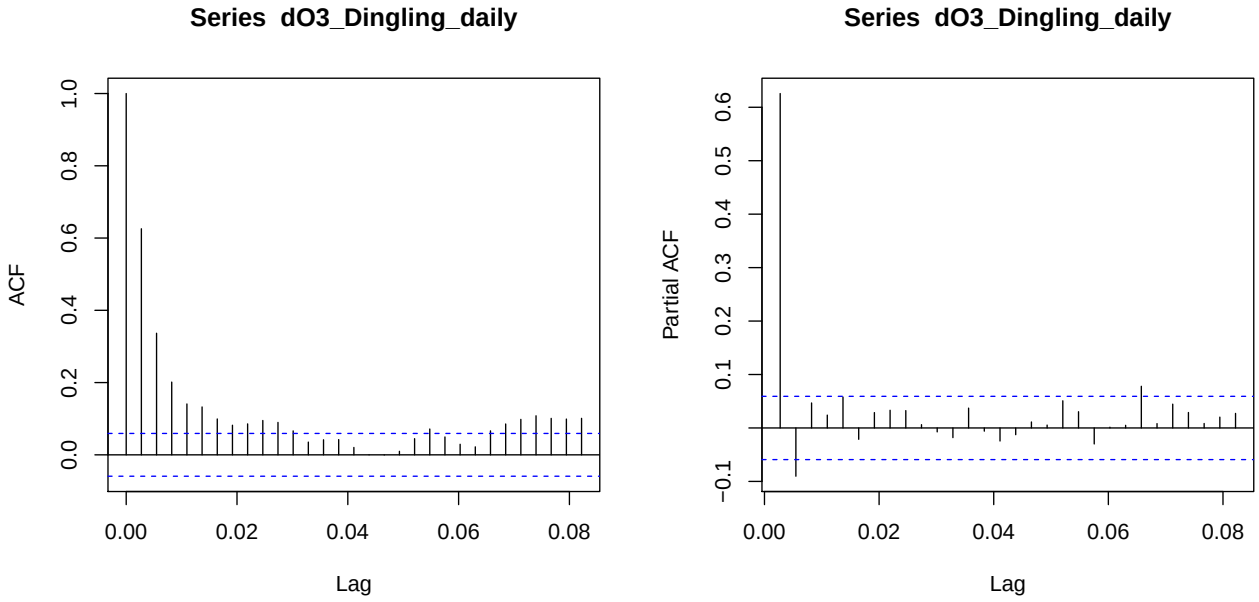
- **Dingling District**

For Dingling District we do the exact same thing as we did for Tiantan District, and we have obtained a stationary process: constant mean around 0, roughly constant variance (with a few very large outliers) and the dependence of observations only depends on lags.



From the ACF plot we can observe the exponentially decaying pattern, and in the PACF plot we can observe two significant peaks.

As a conclusion, a MA(2) process might be correct for ozone daily pollution in Dingling.



4. Modeling: Comparisons of models and Estimations

4.1 Original Hourly Data Set

As we have identified before, we will firstly build a MA(2) process for both Tiantan and Dingling District, afterwards, we will compare it utilizing the AIC criteria with the automatically recognized models by the *auto.arima* function from the *forecast* package. We will not use the *tso* function from the *tsoseries* package due to the fact we could not see any outliers by direct simple inspection and also there are too many observations, therefore it will be very computationally intensive.

We will only consider 24866 observations to train the models (data from 2013 to 2016, 24866/26304=95%), and the rest to test/predict.

- **Tiantan District**

1. MA(2) process by observing the ACF and PACF plots

Which means that the current observation is based on a linear combination of the current and the past 2 innovations that are stochastic with coefficient 1.08 (not invertible) and 0.57:

$$x_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2}$$

$$x_t = a_t + 1.08 a_{t-1} + 0.57 a_{t-2}$$

```
##
## Call:
## arima(x = bt1, order = c(0, 0, 2))
##
## Coefficients:
##          ma1     ma2  intercept
##       1.0785  0.5733     0.0647
## s.e.  0.0054  0.0043     0.0111
##
## sigma^2 estimated as 0.4327:  log likelihood = -24869.71,  aic = 49747.43
```

2. ARIMA(0,1,5) by using function *auto.arima*

By using *auto.arima* function, we got an ARIMA(0,1,5) model which means that the current differenced data is a linear combination of hte past 5 innovations:

$$\nabla x = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} + \theta_4 a_{t-4} + \theta_5 a_{t-5}$$

$$\nabla x = a_t + 0.1345 a_{t-1} - 0.0803 a_{t-2} - 0.1207 a_{t-3} - 0.1094 a_{t-4} - 0.1125 a_{t-5}$$

```
## Series: bt1
## ARIMA(0,1,5)
##
## Coefficients:
##          ma1      ma2      ma3      ma4      ma5
##       0.1345  -0.0803  -0.1207  -0.1094  -0.1125
## s.e.  0.0065   0.0067   0.0076   0.0078   0.0072
##
## sigma^2 estimated as 0.2825:  log likelihood=-19562.4
## AIC=39136.81   AICc=39136.81   BIC=39185.53
```

Conclusions Although the *auto.arima* model (ARIMA(0,1,5)) that has been automatically generated is more complicated and harder to interpret, but the value of AIC has decreased from 49747.43 to 39136.81. So we consider that the second model is better.

- **Dingling District**

1. MA(2) process by observing the ACF and PACF plots

Which means that the current observation is based on a linear combination of current and past 2 innovations that are stochastic with coefficient 1.11 (not invertible) and 0.59:

$$x_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2}$$

$$x_t = a_t + 1.11 a_{t-1} + 0.59 a_{t-2}$$

```
##
## Call:
## arima(x = bt2, order = c(0, 0, 2))
##
## Coefficients:
##          ma1     ma2  intercept
##       1.1139  0.5873     0.0767
## s.e.  0.0055  0.0041     0.0096
##
## sigma^2 estimated as 0.313:  log likelihood = -20841.44,  aic = 41690.88
```

2. ARIMA(0,1,5) by using function *auto.arima*

By using *auto.arima* function, we got an ARIMA(0,1,5) model which means that the current differenced data is a linear combination of the past 5 innovations with the following coefficients:

$$\nabla x = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} + \theta_4 a_{t-4} + \theta_5 a_{t-5}$$

$$\nabla x = a_t + 0.0667 a_{t-1} - 0.0595 a_{t-2} - 0.0496 a_{t-3} - 0.0680 a_{t-4} - 0.0610 a_{t-5}$$

```
## Series: bt2
## ARIMA(0,1,5)
##
## Coefficients:
##          ma1      ma2      ma3      ma4      ma5
##       0.0667  -0.0595  -0.0496  -0.0680  -0.0610
## s.e.  0.0064   0.0064   0.0068   0.0068   0.0066
##
## sigma^2 estimated as 0.1567:  log likelihood=-12234.18
## AIC=24480.36   AICc=24480.37   BIC=24529.09
```

Conclusions

We have the same case as we had before, the value of the AIC has been reduced very significantly, from 41690.88 to 24480.37. Due to this reason, it makes more sense to take the ARIMA(0,1,5) model that the *auto.arima* function has identified.

4.2 Modified Daily Data Set

- **Tiantan District**

By observing the ACF and PACF, it is possible to model an ARIMA(2,0,2) process in the first place.

So we will firstly build an ARIMA(2,0,2) process, and compare it by AIC criteria with the automatically recognized models by using the *auto.arima* function from the *forecast* package and the *tso* function from the *tsoseries* package (although we cannot identify any outliers by simple inspection).

We will only consider 1000 observations to train the models (1000/1096=91%), and the rest to test/predict.

1. ARIMA(2,0,2) process by observing the ACF and PACF plots

Which means that the current observation is based on a linear combination of current and past 2 innovations that are stochastic and past 2 observations with the following coefficients:

$$x = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \phi_1 x_{t-1} + \phi_2 x_{t-2}$$

$$x = a_t - 0.2575 a_{t-1} + 0.0388 a_{t-2} + 0.7969 x_{t-1} - 0.2575 x_{t-2}$$

```
##
## Call:
## arima(x = bt3, order = c(2, 0, 2))
##
## Coefficients:
##          ar1      ar2      ma1     ma2  intercept
##       0.7969  -0.2575  -0.2575  0.0388     1.8167
## s.e.  0.4719   0.1560   0.4718  0.1332     1.5746
##
## sigma^2 estimated as 894.3:  log likelihood = -4995.33,  aic = 10002.65
```

2. ARIMA(3,1,2) by using the *auto.arima* function

By using *auto.arima* function, we got an ARIMA(3,1,2) model which means that the current differenced data is a linear combination of past 2 innovations and past 3 observations with the following coefficients:

$$\nabla x = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3}$$

$$\nabla x = a_t - 1.8126 a_{t-1} + 0.8186 a_{t-2} + 1.3440 x_{t-1} - 0.5296 x_{t-2} + 0.0437 x_{t-3}$$

```
## Series: bt3
## ARIMA(3,1,2)
##
## Coefficients:
##          ar1      ar2     ar3      ma1     ma2
##       1.3440  -0.5296  0.0437  -1.8126  0.8186
## s.e.  0.1098   0.0779  0.0387   0.1048  0.1037
##
## sigma^2 estimated as 890.7:  log likelihood=-4987.09
## AIC=9986.18   AICc=9986.26   BIC=10015.84
```

3. ARIMA(2,1,2) by using the *tso* function

By using *tso* function, we got an ARIMA(2,1,2) model which means that the current differenced data is a linear combination of past 2 innovations and past 2 observations with the following coefficients:

$$\nabla x = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \phi_1 x_{t-1} + \phi_2 x_{t-2}$$

$$\nabla x = a_t - 1.5017 a_{t-1} + 0.5070 a_{t-2} + 1.0152 x_{t-1} - 0.3327 x_{t-2}$$

```
## Series: bt3
## Regression with ARIMA(2,1,2) errors
##
## Coefficients:
##          ar1      ar2      ma1     ma2       TC39     AO137     TC148
##       1.0152  -0.3327  -1.5017  0.5070  -143.5497  106.7970  123.1402
## s.e.  0.1712   0.0773   0.1757  0.1747    27.0440   25.5033   27.0375
##           TC486
##       -110.1152
## s.e.    26.7062
##
## sigma^2 estimated as 827.2:  log likelihood=-4947.57
## AIC=9913.14   AICc=9913.32   BIC=9957.63
##
## Outliers:
##   type ind     time coefhat  tstat
## 1   TC  39  2014:98  -143.5 -5.308
## 2   AO 137 2014:196   106.8  4.188
## 3   TC 148 2014:207   123.1  4.554
## 4   TC 486 2015:180  -110.1 -4.123
```

Conclusions

The models that are automatically fitted using *auto.arima* and *tso* are more complex, and the values of AIC have only decreased from 10002.65 to 9986.26 and 9913.32 respectively, as a consecuence of that, we will pick the first model, ARIMA(2,0,2).

- **Dingling District**

By observing the ACF and PACF, it is possible to model a MA(2) process as for hourly data.

1. MA(2) process by observing the ACF and PACF plots

Which means that the current observation is based on a linear combination of current and past 2 innovations that are stochastic with coefficients 0.6606 and 0.2553:

$$x_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2}$$
$$x_t = a_t + 0.6606 a_{t-1} + 0.2553 a_{t-2}$$

```
##
## Call:
## arima(x = bt4, order = c(0, 0, 2))
##
## Coefficients:
##          ma1     ma2  intercept
##       0.6606  0.2553     5.7051
## s.e.  0.0289  0.0288     2.1077
##
## sigma^2 estimated as 1257:  log likelihood = -5171.78,  aic = 10351.55
```

2. ARIMA(0,1,4) by using the *auto.arima* function

By using the *auto.arima* function, we got an ARIMA(0,1,4) model which means that the current differenced data is a linear combination of past 4 innovations with the following coefficients:

$$\nabla x_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} + \theta_4 a_{t-4}$$
$$\nabla x_t = a_t - 0.3029 a_{t-1} - 0.3413 a_{t-2} - 0.1731 a_{t-3} - 0.1329 a_{t-4}$$

```
## Series: bt4
## ARIMA(0,1,4)
##
## Coefficients:
##           ma1      ma2      ma3      ma4
##       -0.3029  -0.3413  -0.1731  -0.1329
## s.e.   0.0311   0.0315   0.0337   0.0314
##
## sigma^2 estimated as 1226:  log likelihood=-5152.81
## AIC=10315.61   AICc=10315.67   BIC=10340.33
```

3. ARIMA(0,1,3) by using the *tso* function

By using the *auto.arima* function, we got an ARIMA(0,1,3) model which means that the current differenced data is a linear combination of past 3 innovations with the following coefficients:

$$\nabla x_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3}$$

$$\nabla x_t = a_t - 0.3948 a_{t-1} - 0.3674 a_{t-2} - 0.1888 a_{t-3}$$

```
## Series: bt4
## Regression with ARIMA(0,1,3) errors
##
## Coefficients:
##           ma1      ma2      ma3     TC148     TC855
##       -0.3948  -0.3674  -0.1888  150.2383  300.0234
## s.e.   0.0309   0.0361   0.0330   32.0818   33.2037
##
## sigma^2 estimated as 1136:  log likelihood=-5112.97
## AIC=10237.95   AICc=10238.03   BIC=10267.61
##
## Outliers:
##   type ind     time coefhat tstat
## 1   TC 148 2014:207   150.2 4.683
## 2   TC 855 2016:184   300.0 9.036
```

Conclusions:

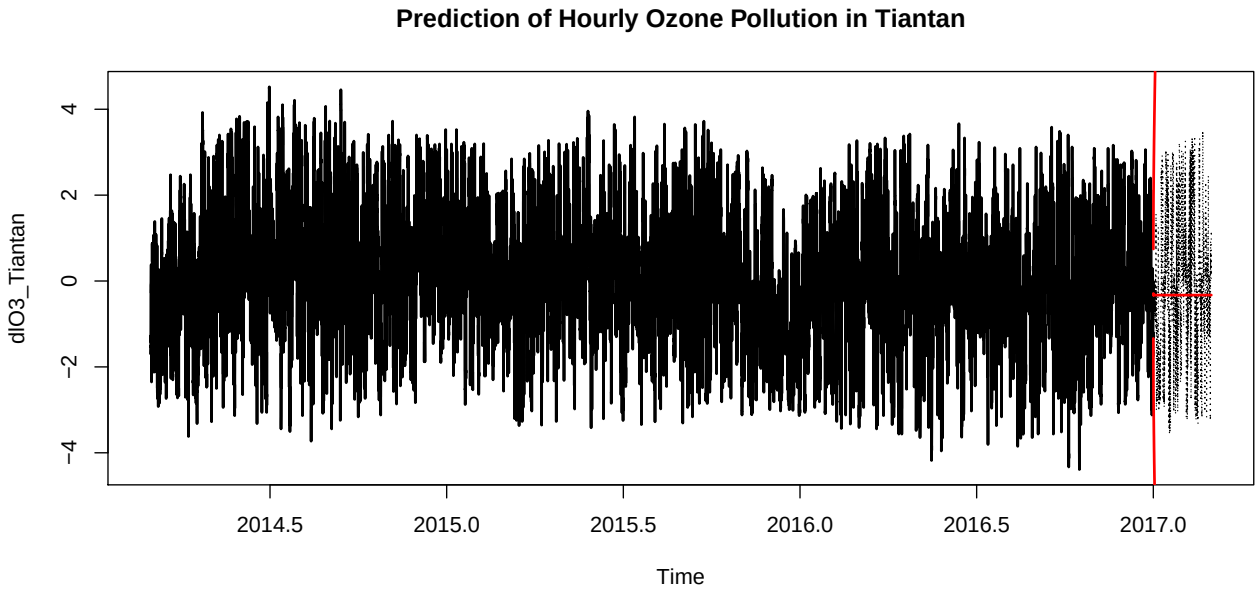
The models identified by the *auto.arima* and *tso* functions are more complex, and the values of the AIC have only decreased very little, from 10351.55 to 10315.67 and 10238.03 respectively, as a consecuence of that, we will pick the first model, ARIMA(0,0,2).

5. Forecast

5.1 Original Hourly Data Set

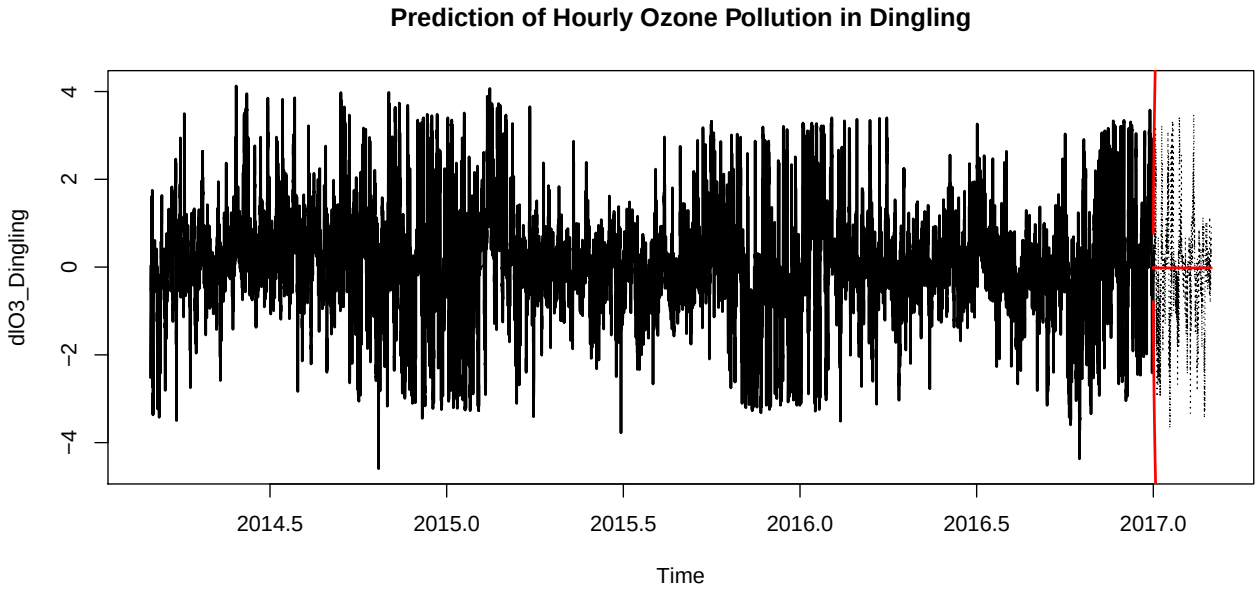
- **Tiantan District**: Model picked: ARIMA(0,1,5)

As the *Hourly Ozone Pollution Series in Tiantan* series is invertible. In this case, its mean converges to 0 after 5 time periods because it contains an MA(5) component, which is hardly visible as we have a large number of observations, and the integrated part doesn't change anything since all the predictions for an integrated process are the last observation. Also, by observing the confidence bounds, it is noticeable that they have diverged to infinity. This is caused by the integrated part, and due to the fact that we are predicting a very large amount of observations (1438), but all the observations are within our bounds.



- **Dingling District** Model picked: ARIMA(0,1,5)

For the case of the *Hourly Ozone Pollution Series in Dingling* TS, we get roughly same conclusions since we have same models and similar processes. Although it is very surprising that the air pollution for both such a centric district and a suburban district are likely the same. The predictions of the process converge to 0 after 5 periods of time given the MA(5) part of the model, and the confidence intervals diverge to infinity because of the integrated part.

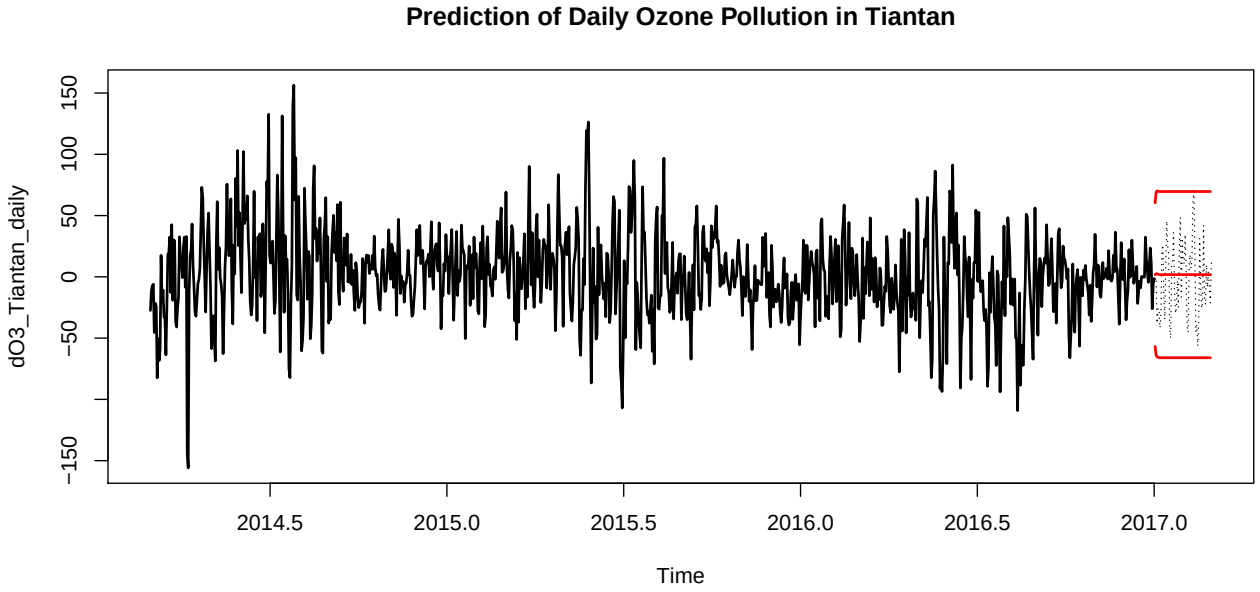

Prediction of Hourly Ozone Pollution in Dingling
dlO3_Dingling
4
2
0
−2
−4
2014.5
2015.0
2015.5
2016.0
2016.5
2017.0
Time

5.2 Modified Daily Data Set

- **Tiantan District** Model picked: ARIMA(2,0,2)

As the *Daily Ozone Pollution Series in Tiantan* TS is a stationary process, the mean and the confidence bounds always converge to a constant. It is converging very fast due to the fact that the coefficients are very small. The first 2 observations are very different from mean due to the moving average part.

Additionaly, we can observe that all the true observations are inside the confidence intervals, although some other observationes observed before are out.



- **Dingling District** Model picked: ARIMA(0,0,2)

As the *Daily Ozone Pollution Series in Dingling* TS is a stationary process, the mean and the confidence bounds always converge to a constant. The first 2 observations are different from the mean due to the moving average part, after those 2 predictions, all others future instances are predicted as the mean of the process.