Final Project for time series–Analysis of AQI of Delhi 2015-2020

wang jiayun

2021/10/29

Contents

Abstract

In this report, I focused on monthly air quality in Delhi, India, from 2015 to 2020.I did an ADF test to make sure our target series was fixed.Then, using the data from 2015 to 2019, SARIMA(0,0,1)×(0,1,1)12 model was fitted based on sample ACF and PACF curves to predict the air quality index for 12 months from August 2019 to August 2020.

By observing the difference between the predicted value and the actual value, it is found that the prediction effect of the model is not good after April 2020, and the actual value is far lower than the predicted value. Combined with the report, it is found that April 2020 is the time point of the outbreak of COVID-19 in India, which indicates that the analysis of data can detect the occurrence of some special events.

1.Description of data

The data set contains air quality data and AQI (Air Quality Index) at daily level of various stations across multiple cities in India.

A glance at the raw data set:

```
##     City      Date  PM2.5   PM10    NO   NO2    NOx    NH3    CO  SO2    O3
## 1  Delhi  2015/1/1 313.22 607.98 69.16 36.39 110.59  33.85 15.20 9.25 41.68
## 2  Delhi  2015/1/2 186.18 269.55 62.09 32.87  88.14  31.83  9.54 6.65 29.97
## 3  Delhi  2015/1/3  87.18 131.90 25.73 30.31  47.95  69.55 10.61 2.65 19.71
## 4  Delhi  2015/1/4 151.84 241.84 25.01 36.91  48.62 130.36 11.54 4.63 25.36
## 5  Delhi  2015/1/5 146.60 219.13 14.01 34.92  38.25 122.88  9.20 3.33 23.20
## 6  Delhi  2015/1/6 149.58 252.10 17.21 37.84  42.46 134.97  9.44 3.66 26.83
## 7  Delhi  2015/1/7 217.87 376.51 26.99 40.15  52.41 134.82  9.78 5.82 28.96
## 8  Delhi  2015/1/8 229.90 360.95 23.34 43.16  51.21 138.13 11.01 3.31 30.51
## 9  Delhi  2015/1/9 201.66 397.43 19.18 38.56  45.60 140.60 11.09 3.48 32.94
## 10 Delhi 2015/1/10 221.02 361.74 24.79 46.39  55.19 134.06  9.70 5.91 34.12
##    Benzene Toluene Xylene AQI AQI_Bucket
## 1    14.36   24.86   9.84 472     Severe
## 2    10.55   20.09   4.29 454     Severe
## 3     3.91   10.23   1.99 143   Moderate
## 4     4.26    9.71   3.34 319  Very Poor
## 5     2.80    6.21   2.96 325  Very Poor
## 6     3.63    7.35   3.47 318  Very Poor
## 7     4.93    9.42   5.21 353  Very Poor
## 8     5.80   11.40   4.83 383  Very Poor
## 9     5.25   11.12   5.26 375  Very Poor
## 10    4.87    9.44   4.76 376  Very Poor
```

It is published by Vopani on Kaggle(https://www.kaggle.com/rohanrao/calculating-aqi-air-quality-index-tutorial/data?scriptVersionId=41199538).The data has been made publicly available by the Central Pollution Control Board: https://cpcb.nic.in/ which is the official portal of Government of India.

AQI is short for air quality index. All the eight pollutants (including PM2.5,PM10,NO,VO__2,NO__x,NH__3,CO,SO__2,O__3) may not be monitored at all the locations. Overall AQI is calculated only if data are available for minimum three pollutants out of which one should necessarily be either PM2.5 or PM10. Else, data are considered insufficient for calculating AQI. Similarly, a minimum of 16 hours' data is considered necessary for calculating sub index.

For the sake of my use, I chose Delhi, one of India's most representative cities, for my analysis. I will aggregate daily data on monthly base and only record the mean of AQI every month. So I may preprocess the raw data set before I do the general time series analysis.

2.The goal of analysis

On March 25 2020, the Indian government placed its population of more than 1.3 billion citizens under lockdown in an effort to reduce the spread of the COVID-19 disease. All non-essential shops, markets and places of worship were closed with only essential services including water, electricity and health services remaining active.

Citizens started to experience better air quality so much so that the scenic Dhauladhar Peaks of Himachal Pradesh became visible from neighbouring states. On normal days, these peaks lie hidden behind he film of smog. Here we have access to a large amount of granular data relating to the concentration of major air pollutants in India and it will be interesting to see if the claim of reduced air pollution is being actually backed by data.

To sum up, the goal of this analysis is to establish a time series model to predict AQI in delhi. I will compare the predict data with actual data to see Whether the epidemic has really affected Delhi's air quality as reported, which will be an interesting insight.

3.Data process and plot

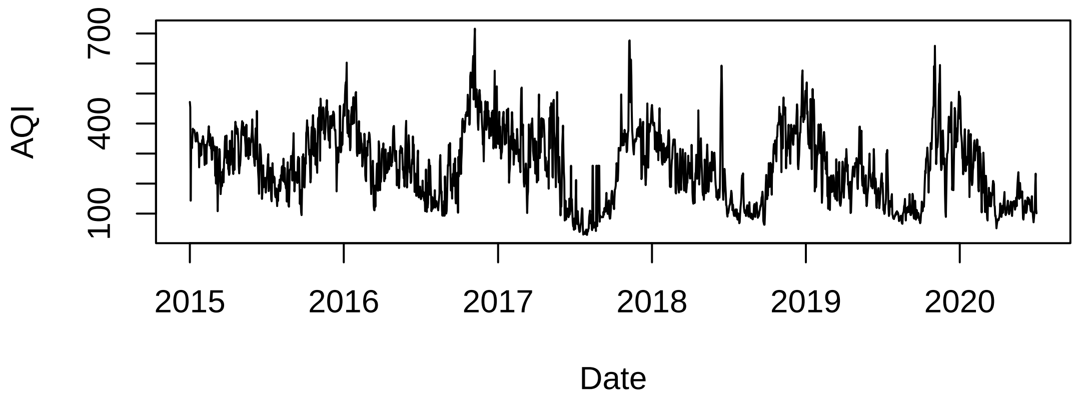
The dimension of original dataset:

```
## [1] 2009   16
```

As to attributes, AQI can represent the level of pollution, so we only focus on the training time series of AQI monthly.Thus I will ignore other attributes. plot of daily data:

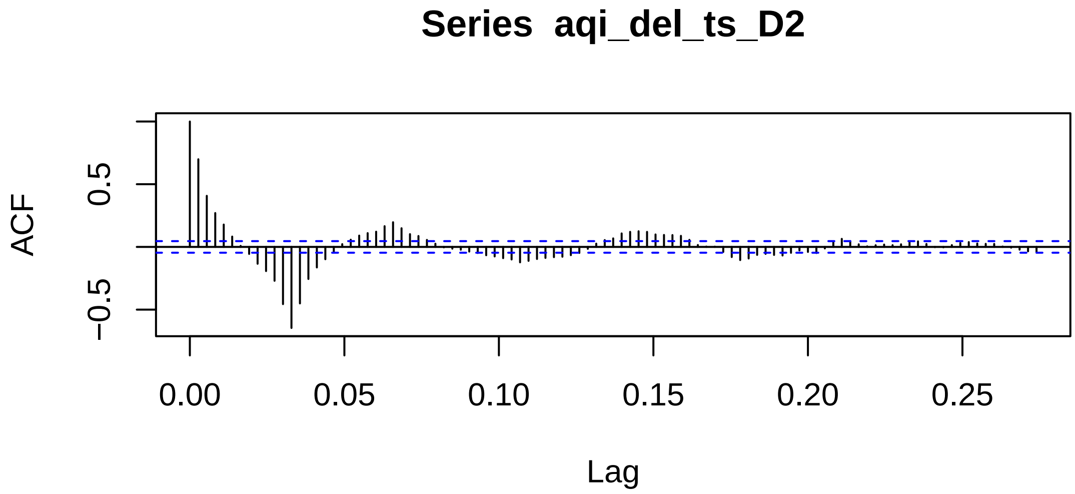
```
##
##  10 values imputed to 259.4877
```

```
##       Date                 AQI
##  Min.   :2015-01-01   Min.   : 29.0
##  1st Qu.:2016-05-17   1st Qu.:162.0
##  Median :2017-10-01   Median :259.0
##  Mean   :2017-10-01   Mean   :259.5
##  3rd Qu.:2019-02-15   3rd Qu.:345.0
##  Max.   :2020-07-01   Max.   :716.0
```



As you can see from the graph, the data is very stable.The reason why we have 2009 observatins is this dataset records the daily AQI since 2015-01-01 to 2020-07-01. I tried to use the data in days for analysis, but after making two seasonal differences, its ACF still showed strong regularity, it is hard for me to handle. So I decided to reduce the data scale for analysis, so I chose to calculate the mean of AQI.

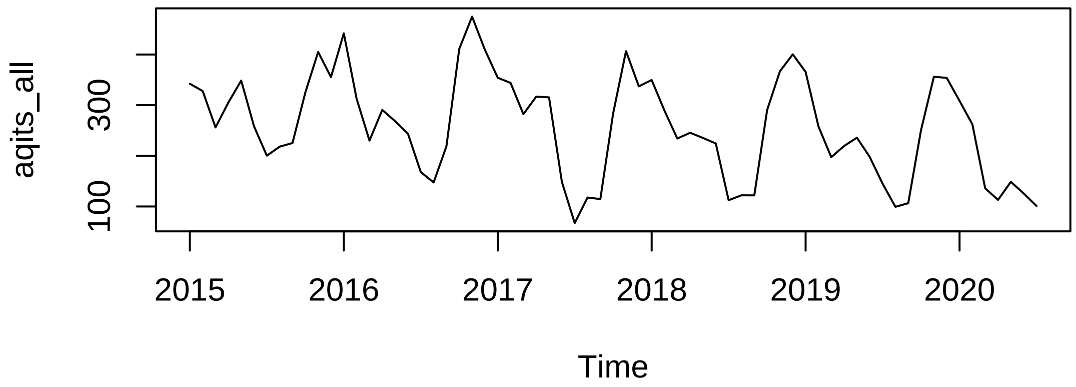
```
## [1] 2009    2
```



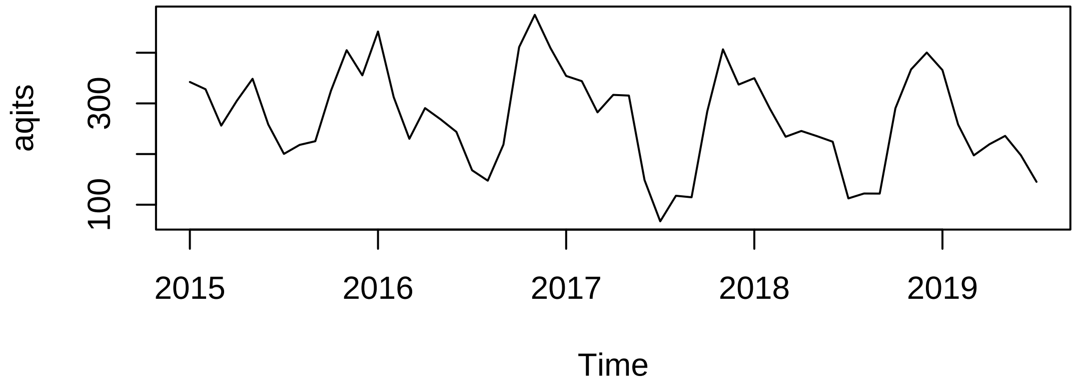
To get reasonable data size, I will calculate the mean of data in each month. Before calculating the monthly data, the missing data values are processed first.Using the summary function, missing values are found in the data set.After checking the number of missing values, it is found that there are 10 missing values in total. In general, 5% of the samples have missing values, which is a relatively safe proportion. Here, the proportion of missing values is 10/2009= 0.2%. So the method of filling the missing values with average values is adopted here.After processing the missing value, convert the data type from data.frame to ts for later processing.

After the preprocess, we will get 67 observations. We will do model fitting on first 55 data points and the last 12 points are leaved as testing data.

The plot of processed monthly data:



The plot of processed monthly data (training):



We can see from the image of monthly data that the data is generally stationary, but there is a strong seasonality. In order to further verify the stability of the data, we use unit root test.

The results from unit root test (ADF test) verifies our observation:

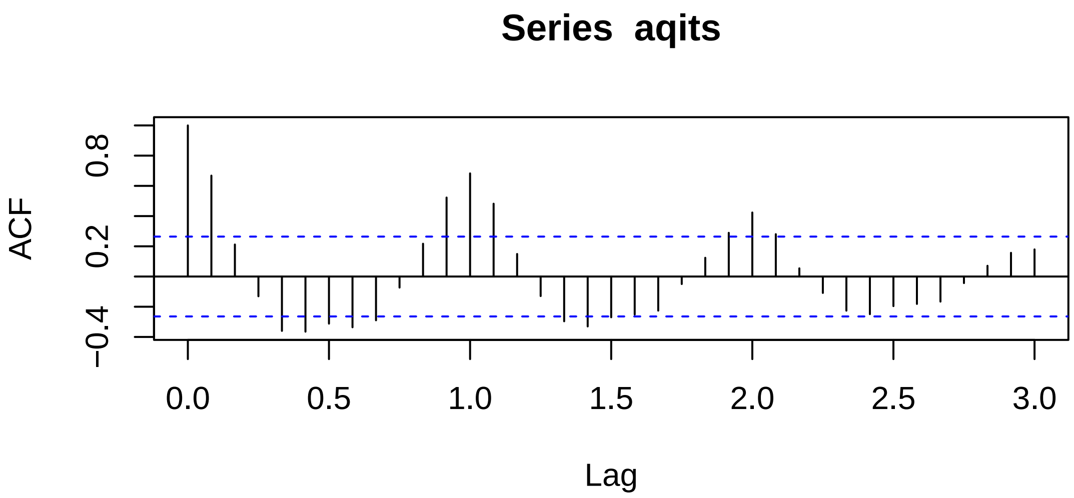
```
## Warning in adf.test(aqits): p-value smaller than printed p-value
```

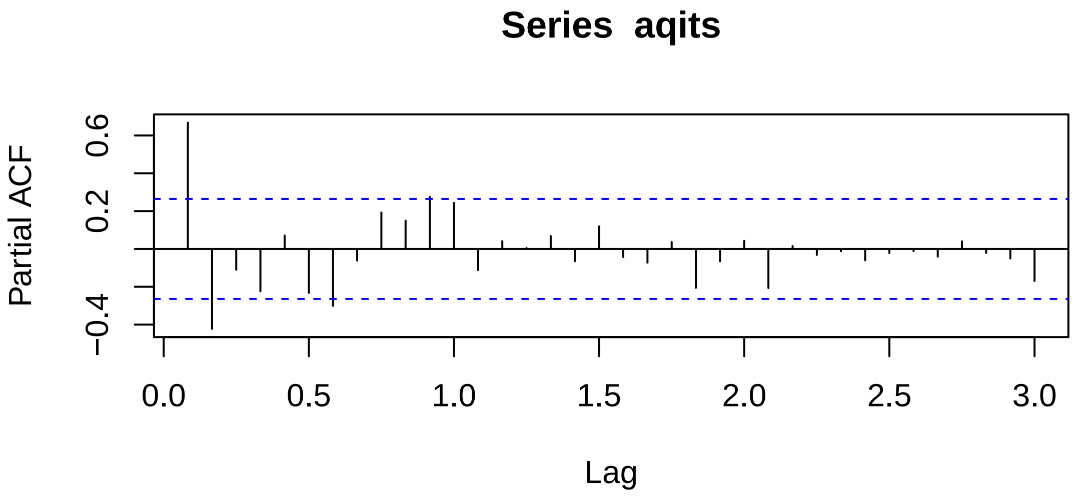
```
##
##  Augmented Dickey-Fuller Test
##
```

```
## data:  aqits
## Dickey-Fuller = -4.5784, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

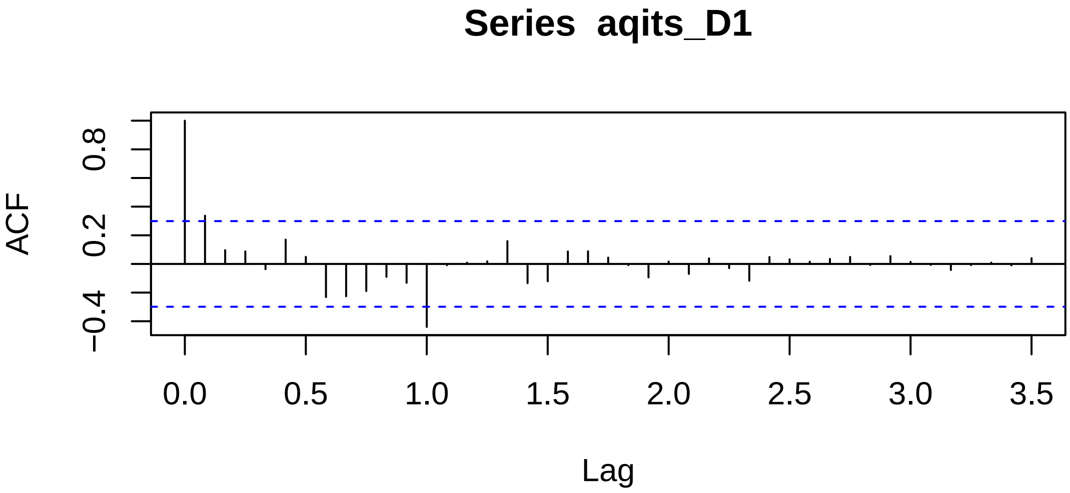
since p-value is extremely small, we reject the null hypothesis and conclude that our Monthly AQI is stationary.

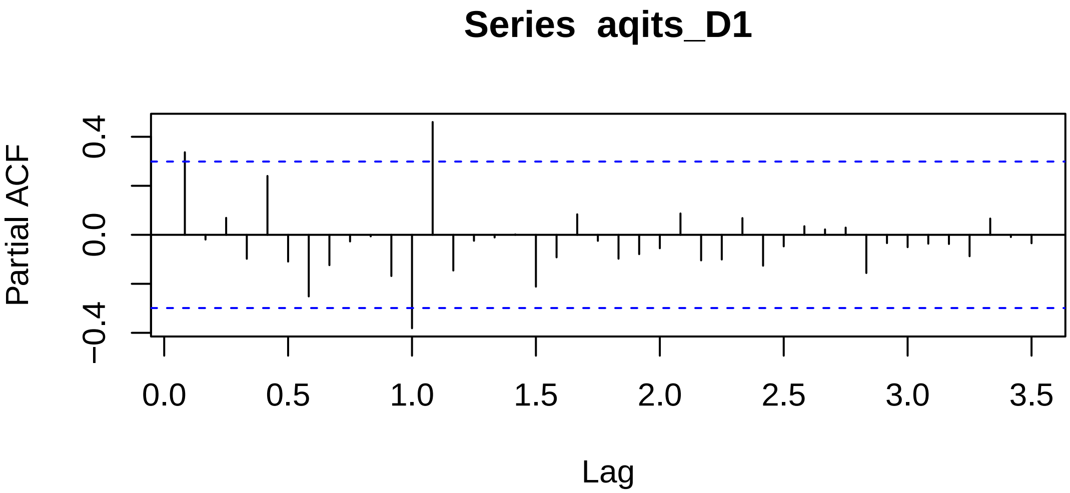
4.Sample ACF and PACF





ACF has certain characteristics, it's very regular, it's a little bit like a sine function, and it decays slowly and decays in such a way that the period is exactly 12. Combined with the periodicity shown by the time sequence diagram, we performed a seasonal difference on the data. According to the prior knowledge of the air pollution and the plot of series, I will try the seasonal ARMA model with one year period(S=12).





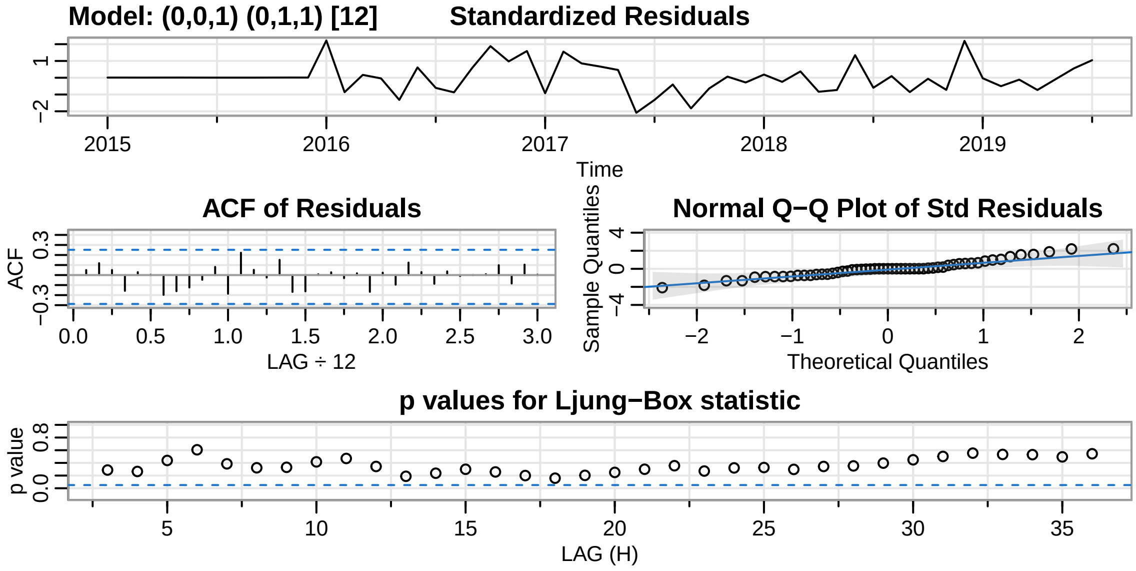
It is found that the data after seasonal difference is still stationary by unit root test. After seasonal difference, we can find that ACF decay quickly, ACF is cut off at a lag 1s, while PACF tails off, These results implies an SMA(1),P=0,Q=1.

For non-seasonal part, both sample ACF and PACF at the lower lags are tailing off. We would first try MA(1,1) within seasons. i.e p=1,q=1.Next part, I will fit an SARIMA model on our time series.

5. Fitting a SARIMA model

Based on the information given in previous parts, it seems that SARIMA model is the most possible model to fit our time series since the whole process demonstrates the periodic pattern and the non-seasonal part shows the dependent pattern.We first try an ARIMA(0,0,1)×(0,1,1)[12]:

```
## initial  value 4.028781
## iter   2 value 3.797915
## iter   3 value 3.766588
## iter   4 value 3.753130
## iter   5 value 3.746468
## iter   6 value 3.743981
## iter   7 value 3.743879
## iter   8 value 3.743878
## iter   8 value 3.743878
## final  value 3.743878
## converged
## initial  value 3.747393
## iter   2 value 3.745541
## iter   3 value 3.741309
## iter   4 value 3.735222
## iter   5 value 3.735096
## iter   6 value 3.735073
## iter   7 value 3.735073
## iter   7 value 3.735073
## iter   7 value 3.735073
## final  value 3.735073
## converged
```



```
## $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  constant
##       0.4390  -1.0000   -1.3560
## s.e.  0.1449   0.3111    0.4103
##
```

```
## sigma^2 estimated as 1144:  log likelihood = -221.62,  aic = 451.24
##
## $degrees_of_freedom
## [1] 40
##
## $ttable
##          Estimate     SE t.value p.value
## ma1         0.439 0.1449  3.0290  0.0043
## sma1       -1.000 0.3111 -3.2144  0.0026
## constant   -1.356 0.4103 -3.3050  0.0020
##
## $AIC
## [1] 10.49407
##
## $AICc
## [1] 10.50838
##
## $BIC
## [1] 10.6579
```

The p.values of all coefficient are smaller than 0.05, which means all coefficients are significant, which implies that our model describes the date well.

From those model diagnostics, p-values for Ljung-Box statistics imply that the model assumptions are satisfied,since all of them are non-significant. The plot of sample ACF implies no severe autocorrelation among residuals.so the model should be adjusted. We can also see that the Q-Q plot and plot of standardized residuals imply that the normality of standardized residuals is satisfied.

To sum up, our log-transformed ARIMA(0,0,1)×(0,1,1)[12] is adequate to describe the series.

The fitted SARIMA(0,0,1)×(0,1,1)[12]model would be (without constant term):

$$\begin{gathered}
\nabla_{12} x_t = \Theta(B^{12})\theta(B)w_t \\
(1-B^{12})x_t = (1+\Theta B^{12})(1+\theta B)w_t \\
x_t = x_{t-12} + w_t + \theta w_{t-1} + \Theta w_{t-12} + \Theta\theta w_{t-13} \\
x_t = x_{t-12} + w_t + 0.439 w_{t-1} - w_{t-12} - 0.439 w_{t-13}
\end{gathered}$$

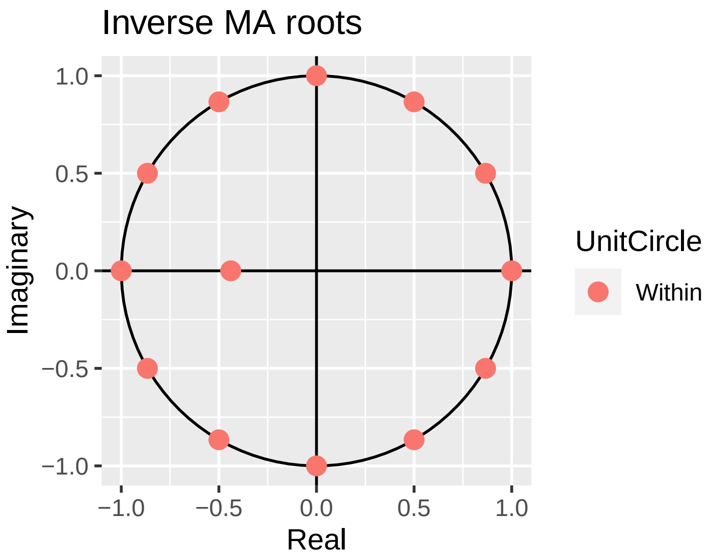
auto.arima returns best ARIMA model according to either AIC, AICc or BIC value. The function conducts a search over possible model within the order constraints provided.

```
##
##  ARIMA(2,0,2)(1,1,1)[12] with drift         : Inf
##  ARIMA(0,0,0)(0,1,0)[12] with drift         : 476.0263
##  ARIMA(1,0,0)(1,1,0)[12] with drift         : 461.9031
##  ARIMA(0,0,1)(0,1,1)[12] with drift         : Inf
##  ARIMA(0,0,0)(0,1,0)[12]                    : 475.2358
##  ARIMA(1,0,0)(0,1,0)[12] with drift         : 474.0193
##  ARIMA(1,0,0)(1,1,1)[12] with drift         : Inf
##  ARIMA(1,0,0)(0,1,1)[12] with drift         : Inf
##  ARIMA(0,0,0)(1,1,0)[12] with drift         : 467.8521
##  ARIMA(2,0,0)(1,1,0)[12] with drift         : 463.1824
##  ARIMA(1,0,1)(1,1,0)[12] with drift         : 463.0998
##  ARIMA(0,0,1)(1,1,0)[12] with drift         : 459.5649
##  ARIMA(0,0,1)(0,1,0)[12] with drift         : 473.7326
##  ARIMA(0,0,1)(1,1,1)[12] with drift         : Inf
##  ARIMA(0,0,2)(1,1,0)[12] with drift         : 463.1254
##  ARIMA(1,0,2)(1,1,0)[12] with drift         : 466.8599
##  ARIMA(0,0,1)(1,1,0)[12]                    : 462.4766
##
##  Best model: ARIMA(0,0,1)(1,1,0)[12] with drift
```

we can see that ARIMA(0,0,1)(1,1,0)[12] with drift is recommended, but our model is signed with inf, which is strange.

The article“Why doesn’t auto.arima() return the model with the lowest AICc value?”may help us.https://robjhyndman.com/hyndsight/badarima/.

The auto.arima() function does not simply find the model with the lowest AICc value. It also carries out several checks to ensure the model is numerically well-behaved.While the Arima() function will never return a model with roots inside the unit circle, the auto.arima() function is even stricter and will not select a model with roots close to the unit circle either. The ARIMA(0,0,1)(0,1,1)[12] model fitted above has roots almost on the unit circle.

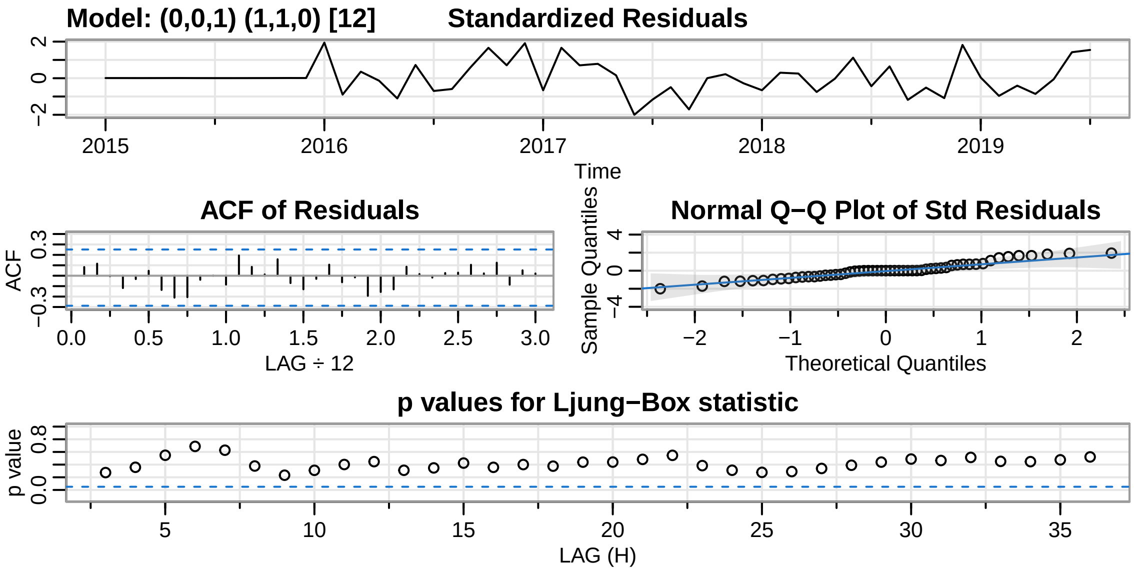


The ARIMA(0,0,1)(0,1,1)[12] model fitted above has roots almost on the unit circle.

Consequently, this model is rejected by auto.arima() because the forecasts will be numerically unstable, and the AICc value is set to Inf to prevent it being selected.

To make our forecast more accurate, i follow the model that R recommends.

```
## initial  value 4.001162
## iter   2 value 3.724912
## iter   3 value 3.639848
## iter   4 value 3.603755
## iter   5 value 3.600826
## iter   6 value 3.595570
## iter   7 value 3.583940
## iter   8 value 3.579783
## iter   9 value 3.579614
## iter  10 value 3.579574
## iter  11 value 3.579566
## iter  12 value 3.579561
## iter  13 value 3.579559
## iter  14 value 3.579559
## iter  14 value 3.579559
## iter  14 value 3.579559
## final  value 3.579559
## converged
## initial  value 3.765625
## iter   2 value 3.754650
## iter   3 value 3.750723
## iter   4 value 3.750081
## iter   5 value 3.749903
## iter   6 value 3.749900
## iter   7 value 3.749899
## iter   7 value 3.749899
## iter   7 value 3.749899
## final  value 3.749899
## converged
```



```
## $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     sar1  constant
##       0.5926  -0.6630   -1.5264
## s.e.  0.1256   0.1143    0.5358
##
## sigma^2 estimated as 1522:  log likelihood = -222.26,  aic = 452.52
```

```
##
## $degrees_of_freedom
## [1] 40
##
## $ttable
##          Estimate     SE t.value p.value
## ma1        0.5926 0.1256  4.7168  0.0000
## sar1      -0.6630 0.1143 -5.8026  0.0000
## constant  -1.5264 0.5358 -2.8486  0.0069
##
## $AIC
## [1] 10.52372
##
## $AICc
## [1] 10.53803
##
## $BIC
## [1] 10.68755
```

It performs well in diagnostic checking.

Next check whether Garch effect exists.

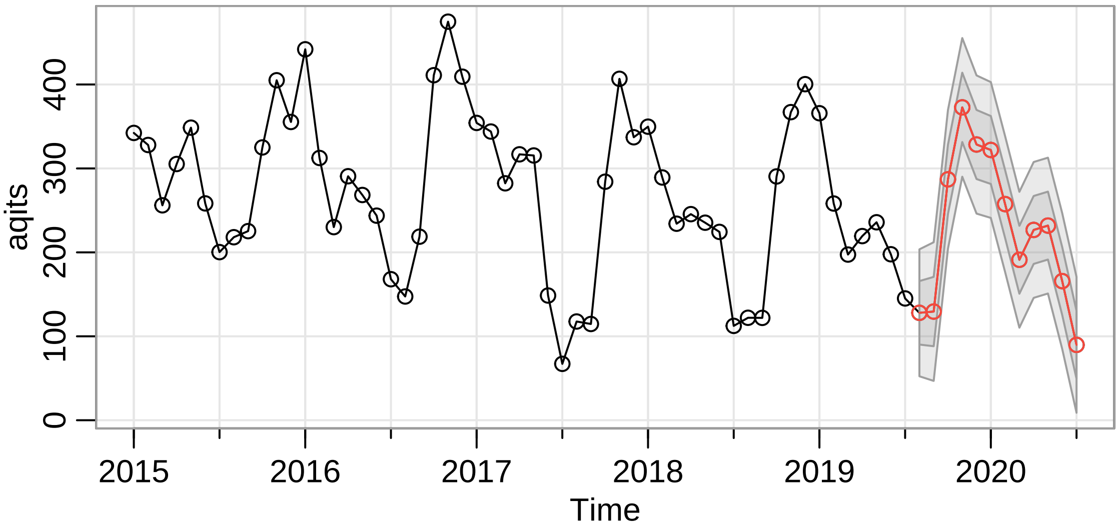
McLeod.Li.test is used to test for conditional heteroscedascity (ARCH). It can be seen from the test results that all points are significantly greater than the significance level of 5%. In other words, the assumption that there is no ARCH effect is accepted, which indicates that SARIMA model has been able to explain the data well.

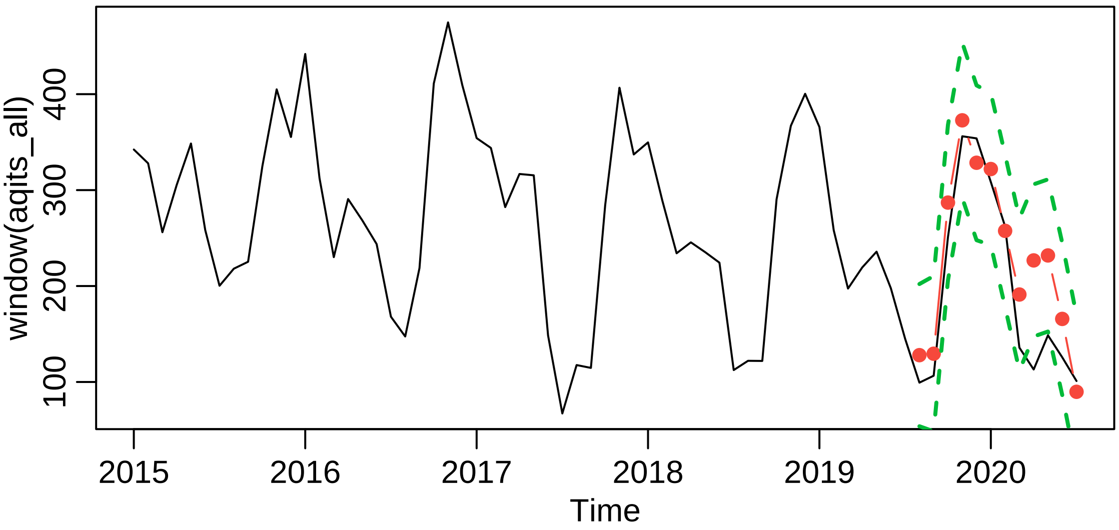
The fitted SARIMA(0,0,1)×(0,1,1)[12]model would be (without constant term):

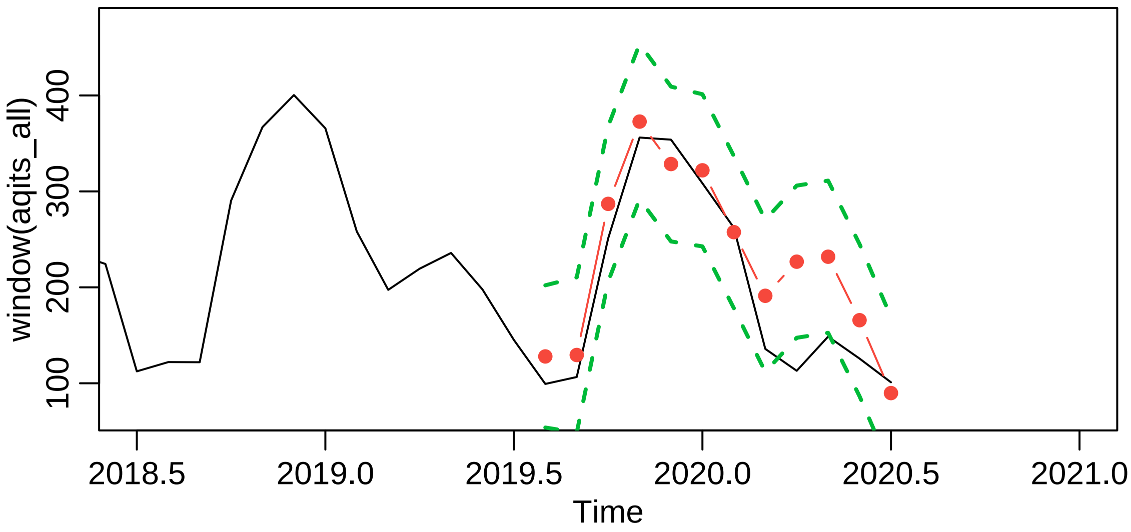
$$\nabla_{12}x_t = \Theta(B^{12})\theta(B)w_t$$
$$(1-B^{12})x_t = (1+\Theta B^{12})(1+\theta B)w_t$$
$$x_t = x_{t-12} + w_t + \theta w_{t-1} + \Theta w_{t-12} + \Theta\theta w_{t-13}$$
$$x_t = x_{t-12} + w_t + 0.5926w_{t-1} - 0.6630w_{t-12} - 0.3929w_{t-13}$$

6. Forecast

The obtained model is used to verify the data of the test set to observe the fitness of the model.







The black broken line in the figure represents the actual AQI data, the red broken line represents the AQI data predicted by the obtained model, and the area between the green broken line represents the 95% confidence interval.

Combined with the table data, it can be seen from the figure that the predicted data of the model from August 2019 to April 2020 is very close to the real data, but the predicted data from May 2020 to August 2020 is significantly lower than the real data.This suggests that a special event may have occurred during the 4 months, leading to an abnormal decrease in AQI.

We all know that COVID-19 broke out in 2020, so we guess that this unusual decline is related to COVID-19. The following is reported by The Guardian,this report can explain why the predicted value is much higher than the actual value:

On March 25 2020, the Indian government placed its population of more than 1.3 billion citizens under lockdown in an effort to reduce the spread of the COVID-19 disease. All non-essential shops, markets and places of worship were closed with only essential services including water, electricity and health services remaining active. Citizens started to experience better air quality so much so that the scenic Dhauladhar Peaks of Himachal Pradesh became visible from neighbouring states. On normal days, these peaks lie hidden behind he film of smog.

7. Conclusion

A careful validation must be done such as the cross validation for time series before we give the robust prediction.As to the pattern capturing, our SARIMA model is adequate to capture the periodicity of the data.

The above analysis shows that The AQI in Delhi has a strong seasonality. The AQI in Delhi increases from August to December, and then gradually decreases. Autumn and winter are the most serious pollution periods.We can use SARIMA model to predict Delhi's air quality very well. If the predicted air quality is far from the actual air quality in a certain period, we can determine that special events may have occurred and take response measures. Such prediction can help us identify abnormal events.

References

Shumway, & Stoffer, R. H. 2017.Time Series Analysis and Its Applications: With R Examples. SpringerInternational Publishing. https://doi.org/10.1007/978-3-319-52452-8.

VOPANI, Air Quality Data in India (2015 - 2020). https://www.kaggle.com/rohanrao/calculating-aqi-air-quality-index-tutorial/data?scriptVersionId=41199538

'It's positively alpine!': Disbelief in big cities as air pollution falls. https://www.theguardian.com/environment/2020/apr/11/positively-alpine-disbelief-air-pollution-falls-lockdown-coronavirus