George Washington University

Predicting the Air Quality Index (AQI) of Shanghai using Weather Features

Team member: Yu Luo, Jason Liu, Ting Huang, Yuwen Luo

DNSC 6219 Time-Series Forecasting

Professor: Refik Soyer

Date: 10 May. 2021

Contents

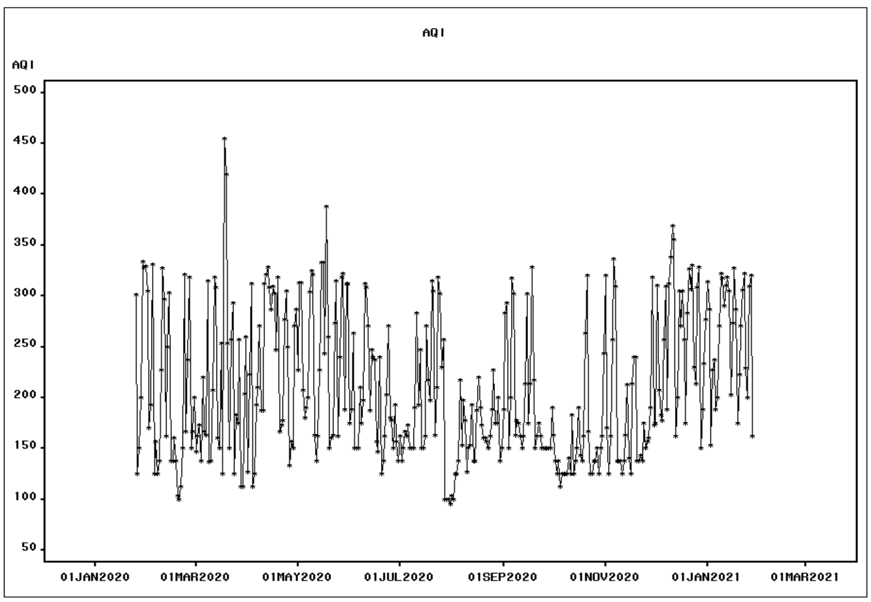
1. Introduction and Overview

1.1 Explanation of Dataset

In this project we will try to predict the AQI in Shanghai using the temperature dataset. This dataset contains about three years' daily records from 01/01/2018 to 01/25/2021. We will also search which weather condition has a higher effect on air quality. We will use Shanghai as the base city, since the air quality of Shanghai has seasonal air conditions, close to sea and has low air quality.

2. Univariate Time-series Models

Since this dataset contains three-year observations, for better visualization purpose, only the most latest one year, from 25 Jan 2020 to 25 Jan 2021 AQI is shown in Fig. 1. The AQI does not seem to have a stable mean. But ACF should be tested to decide whether this is stationary or not.

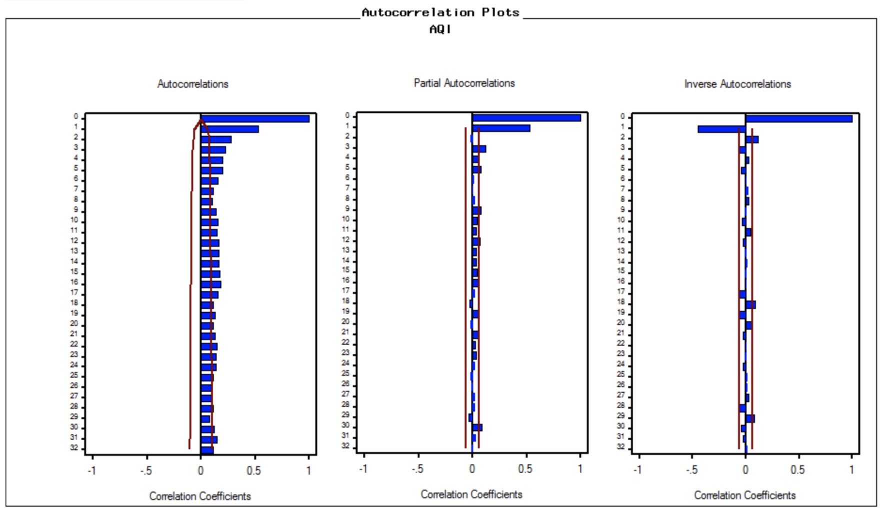


(Figure 1)

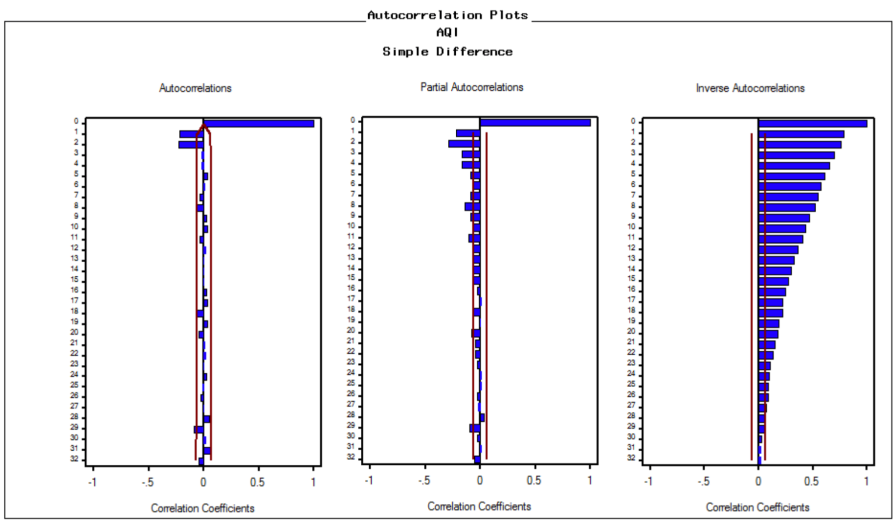
The autocorrelation (fig. 2) has been decreasing at a very slow pace. Therefore, it indicates that the series is not stationary. The first difference is taken as transformation.

The autocorrelation plot of the differenced series (fig. 3) suggests that the differenced data is stationary because the ACF now cuts off at lag 2 and converges to zero after that.

Based on the previous conclusion, we think we could move on to test its linearity and seasonality.



(Figure 2)



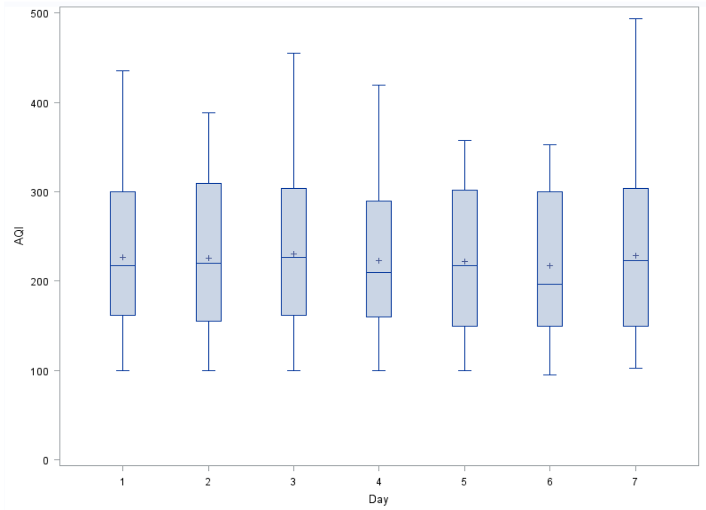
(Figure 3)

2.1 Deterministic Time Series Models and Error Model

Before modeling, the most lastest 150 observations are selected as the validation set which is about 13% of the original time series.

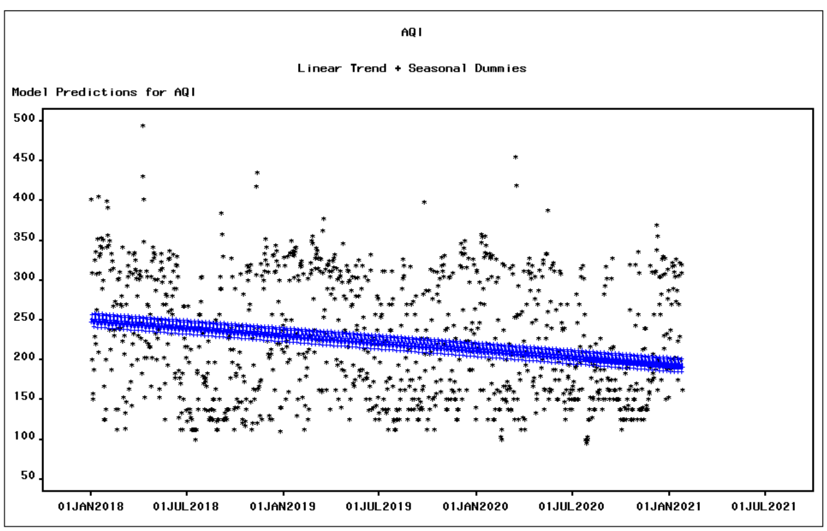
2.1.1 Seasonal Dummies Model

First, boxplot (fig. 4) is created to visualize the daily distribution of AQI. The average of AQI does not seem to be greatly different from each day. Saturday seems to have the lowest average. Some days seem to be skewed distribution such as Sunday, Monday, Wednesday. Therefore, seasonal dummies are still worth investigation.



(Figure 4)

If linearity and weekly seasonal dummies are used into the model, it is statistically significantly different from 0 (fig. 4). While for weekly dummies with Sunday as reference, only one of them is significant while the rest are not. From the plot (fig. 5), directly fit the linear trend and the seasonal dummies do not seem to offer a good prediction. From the residual ACF (fig. 7), there is a slowly decaying trend which means that the residual is not white noise. Therefore, further error models can not be directly fit in. This further implies that the deterministic model using seasonal dummies is not appropriate in this case.



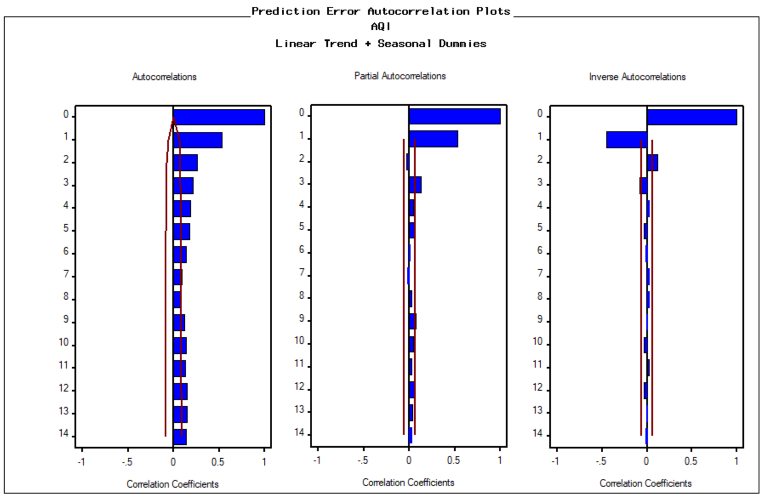
(Figure 5)

Parameter Estimates
AQI
Linear Trend + Seasonal Dummies

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	238.19912	7.7460	30.7512	<.0001
Linear Trend	-0.04310	0.0087	-4.9406	<.0001
Seasonal Dummy 1	18.36918	9.1723	2.0027	0.0471
Seasonal Dummy 2	6.43697	9.1558	0.7031	0.4832
Seasonal Dummy 3	16.35777	9.1558	1.7866	0.0761
Seasonal Dummy 4	11.92604	9.1558	1.3026	0.1948
Seasonal Dummy 5	7.55187	9.1558	0.8248	0.4109
Seasonal Dummy 6	7.40073	9.1558	0.8083	0.4203
Model Variance (sigma squared)	5805	.	.	.

Fit Range: 01JAN2018 to 28AUG2020

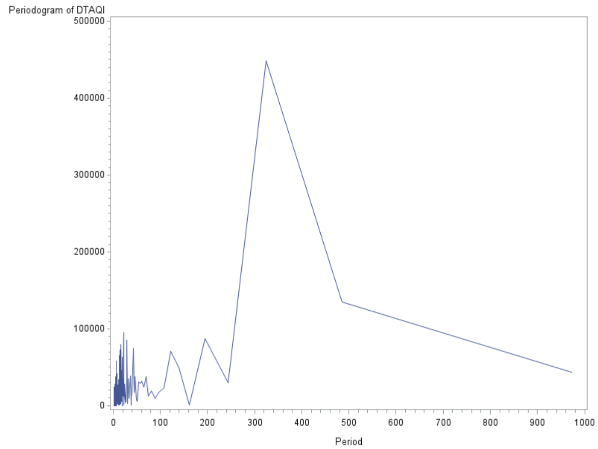
(Figure 6)



(Figure 7)

2.1.2 Cyclical Trend Model

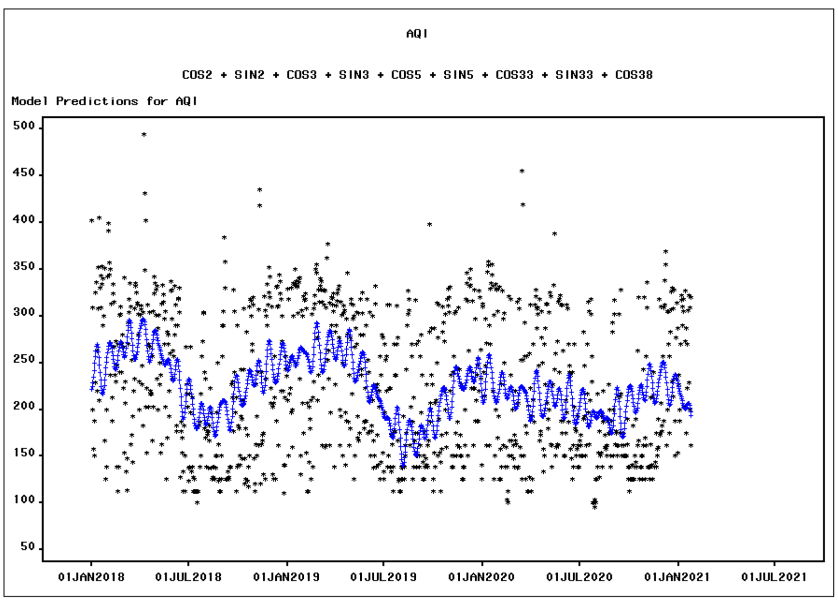
Based on the periodogram and its P-value (fig. 8), the top eight harmonic i with the largest value of periodogram value seems to have p-value larger than 6%. Therefore, the top 8 periods with i = {3,2,45,5,33,52,38,64} are chosen as the potential harmonics to fit in the model.



Obs	FREQ	PERIOD	P_01	RANKING
1	0.01681	373.67	0.74713	1
2	0.03363	186.83	0.12211	2
3	0.37553	16.73	0.10805	3
4	0.49324	12.74	0.09265	4
5	0.05044	124.56	0.08856	5
6	0.29146	21.56	0.08699	6
7	0.21299	29.50	0.07865	7
8	0.35872	17.52	0.07239	8
9	0.01121	560.50	0.06507	9
10	0.14012	44.84	0.05219	10
11	0.17375	36.16	0.05179	11
12	0.11210	56.05	0.05126	12
13	0.30267	20.76	0.04947	13
14	0.87998	7.14	0.04934	14
15	0.10649	59.00	0.04917	15

(Figure 8)

After fitting the cyclical trend, the predicted vs actual plot is shown in fig. 9. The overall trend seems to be modeled, but still most estimated seems to be far from its actual. In fig. 10, some pairs do not have p-values smaller than 0.05 such as sin38 and cos38, sin52 and cos52, sin64 and cos64.



(Figure 9)

Parameter Estimates

AQI

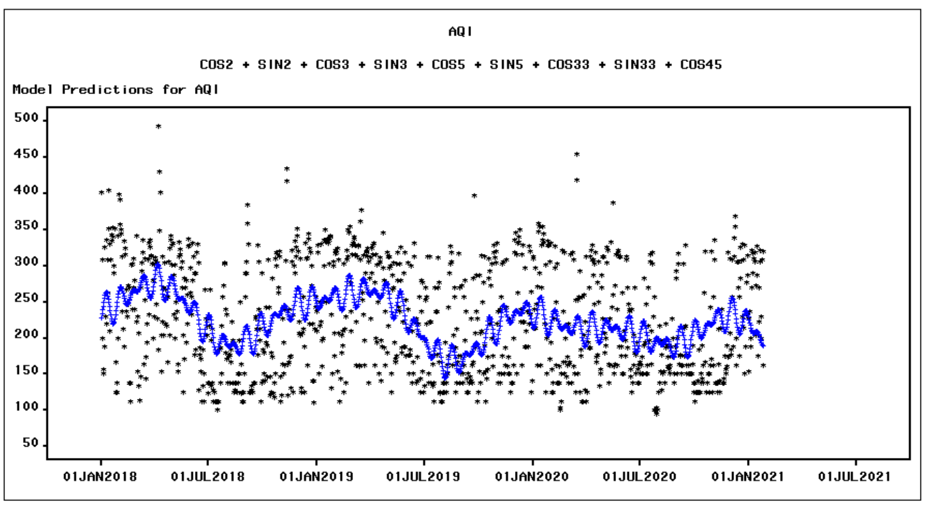
+ COS3 + SIN3 + COS5 + SIN5 + COS33 + SIN33 + COS38 + SIN38 + COS45 + SIN45 + COS52 + SIN52 + COS64 + SIN64 +

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	248.88177	4.7528	52.3648	<.0001
COS2	10.35124	3.1130	3.3252	0.0013
SIN2	-12.99628	3.3778	-3.8476	0.0002
COS3	-7.23405	3.1245	-2.3152	0.0231
SIN3	27.49399	3.2307	8.5102	<.0001
COS5	-13.93612	3.1437	-4.4330	<.0001
SIN5	-4.85994	3.1397	-1.5479	0.1255
COS33	-0.23608	3.1225	-0.0756	0.9399
SIN33	11.77506	3.1263	3.7665	0.0003
COS38	-1.41534	3.1286	-0.4524	0.6522
SIN38	-0.84796	3.1238	-0.2715	0.7867
COS45	-12.63858	3.1276	-4.0410	0.0001
SIN45	2.75712	3.1236	0.8827	0.3800
COS52	0.48158	3.1221	0.1542	0.8778
SIN52	-4.99114	3.1254	-1.5970	0.1141
COS64	-2.94256	3.1222	-0.9425	0.3487
SIN64	-4.22579	3.1202	-1.3543	0.1793
Linear Trend	-0.04770	0.0082	-5.7982	<.0001
Model Variance (sigma squared)	4969	.	.	.

Fit Range: 01JAN2018 to 17OCT2020

(Figure 10)

After removing the non-significant pairs of cyclical trends, the predicted vs actual does not seem that different from the previous plot (fig. 11). All of the pairs now have significant p-values (fig. 12), so does the linear trend. The ACF of residual (fig.13) shows a slowly decaying trend and its PACF cuts off at lag 1. Therefore, AR(1) should be fit in the residual model.



(Figure 11)

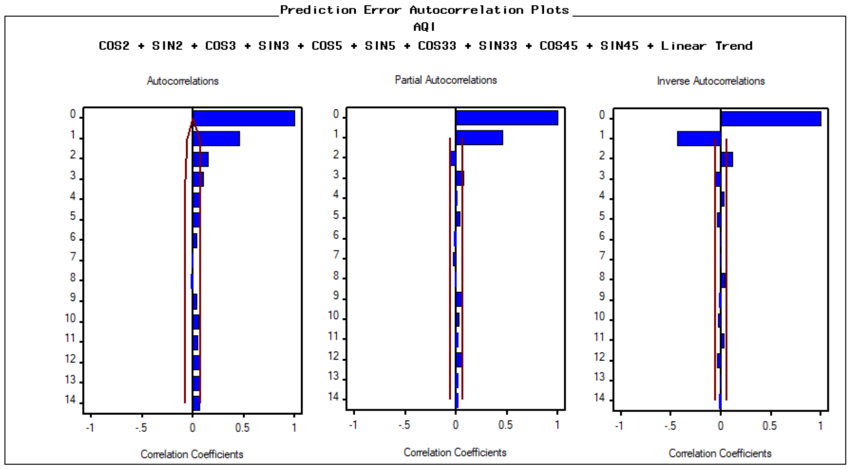
Parameter Estimates

AQI

COS2 + SIN2 + COS3 + SIN3 + COS5 + SIN5 + COS33 + SIN33 + COS45 + SIN45 + Linear Trend

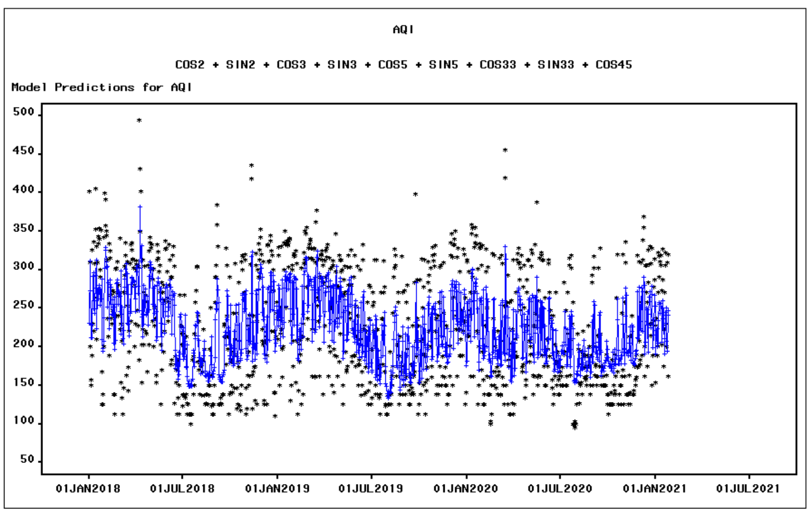
Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	248.78903	4.7502	52.3746	<.0001
COS2	10.28794	3.1120	3.3059	0.0014
SIN2	-12.98868	3.3766	-3.8467	0.0002
COS3	-7.29236	3.1236	-2.3346	0.0218
SIN3	27.49156	3.2297	8.5122	<.0001
COS5	-13.98156	3.1427	-4.4489	<.0001
SIN5	-4.86880	3.1387	-1.5512	0.1244
COS33	-0.34545	3.1187	-0.1108	0.9121
SIN33	11.83668	3.1228	3.7905	0.0003
COS45	-12.80537	3.1224	-4.1011	<.0001
SIN45	2.59087	3.1182	0.8309	0.4083
Linear Trend	-0.04758	0.0082	-5.7872	<.0001
Model Variance (sigma squared)	4966	.	.	.

(Figure 12)

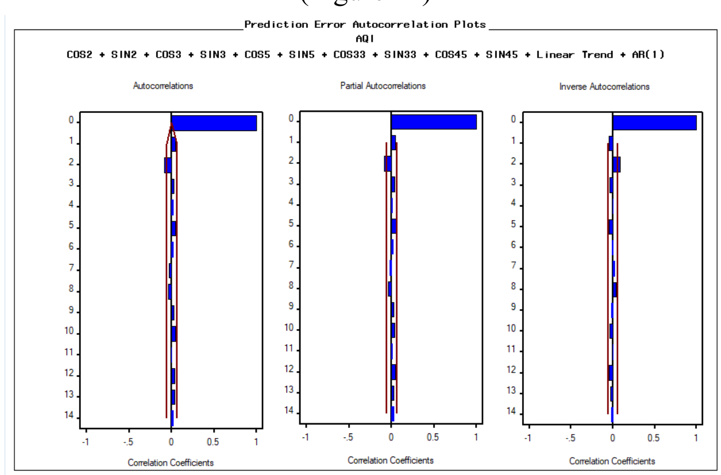


(Figure 13)

After modeling the residual with AR(1), the predicted vs actual (fig. 14) now seems better than the previous. All the parameters are still significant. The predicted values now are much closer than the actual. The ACF (fig. 15) now seems to be a white noise.



(Figure 14)

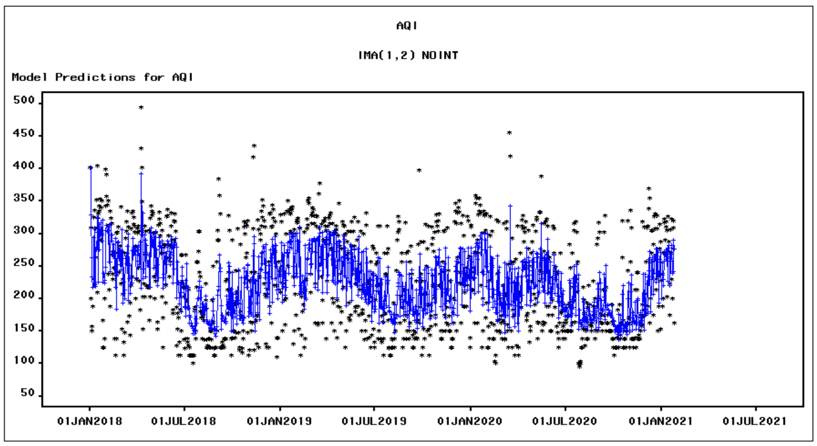


(Figure 15)

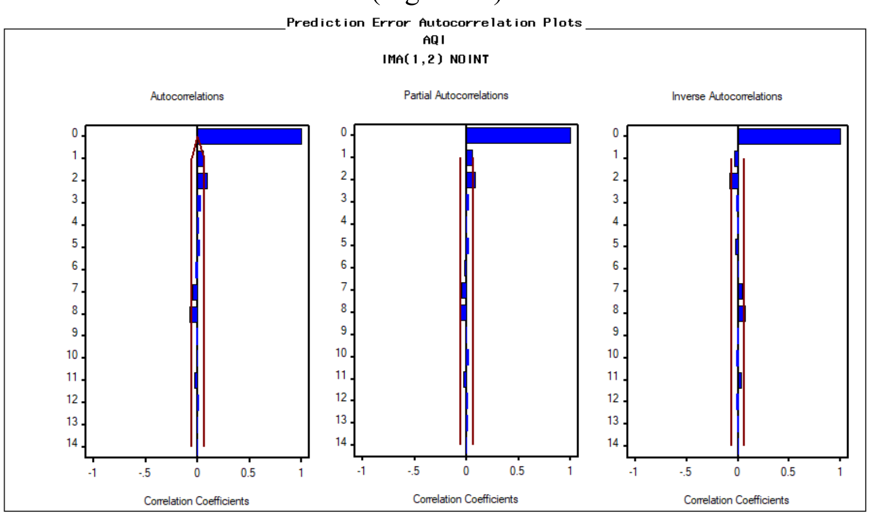
2.2 ARIMA Models

From the conclusion of section 1, AQI is a non-stationary model based on fig. 2. After taking the first difference, ACF (fig. 3) seems to cut off at lag 2. Therefore, MA(2) might be applicable (fig. 16). Furthermore, in fig. 17, no further lag larger than 2 seems to be significant. Therefore, seasonality might not be suitable in this case.

After fitting MA(2) into the first difference, the prediction seems to catch some pattern of the actual. From the ACF, the residual seems to be white noise. Both parameters for MA also seem to be significant (fig. 18).



(Figure 16)



(Figure 17)

Parameter Estimates
AQI
IMA(1,2) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.52170	0.0294	17.7160	<.0001
Moving Average, Lag 2	0.40237	0.0294	13.6706	<.0001
Model Variance (sigma squared)	4245	.	.	.

Fit Range: 01JAN2018 to 28AUG2020

(Figure 18)

Considering to better forecast the extreme value, it has been noticed some holidays in Shanghai have high AQI. Therefore, a new predictor, named 'holiday', was created to indicate whether AQI was made on a holiday or not.

The parameter estimation (fig. 19) has shown that this newly created variable, 'holiday' is not significant. Therefore, 'holiday' should not be included in the model. Neither should intercept since it has p-value larger than 0.05. Therefore, the previous model with MA(2) on first difference should be preferred.

Parameter Estimates
AQI
Holiday + IMA(1,2)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	-0.11180	0.1526	-0.7326	0.4656
Moving Average, Lag 1	0.51960	0.0287	18.0917	<.0001
Moving Average, Lag 2	0.40604	0.0287	14.1419	<.0001
Holiday	7.73349	25.6432	0.3016	0.7636
Model Variance (sigma squared)	4151	.	.	.

Fit Range: 01JAN2018 to 17OCT2020

(Figure 19)

2.3 Models Comparison

From the previous analysis, three models (fig. 20) have been obtained currently. Mean absolute percentage error (MAPE) was chosen to compare the models in the validation set from Oct 18th 2020 to Jan 25th 2021. The ARIMA model has lower MAPE compared with the cyclical trend model.

In terms of model variance, the ARIMA model has model variance 4245 (fig. 18) and the cyclical model has model variance 4966 which is slightly higher. (fig. 12)

Data Set: WORK.NEW Interval: DAY
Series: AQI Browse...
Data Range: 01JAN2018 to 25JAN2021
Fit Range: 01JAN2018 to 17OCT2020
Evaluation Range: 18OCT2020 to 25JAN2021 Set Ranges...

Forecast Model	Model Title	Mean Absolute Percent Error
	COS2 + SIN2 + SIN3 + COS3 + COS5 + SIN5 + COS33 + SIN33 + COS45	26.98427
✓	IMA(1,2) NOINT	21.31070

(Figure 20)

3. Multivariate Time Series Models

Among the whole dataset, the following three indicators are chosen to fit in the multivariate models:

- PM 2.5 - One of the most minuscule categories of these airborne hazards are called particulate matter (PM) 2.5
- PM 10 - Organic particles, or particulate matter, as in smoke, measuring between 2.5 and 10 microns in diameter.
- Carbon Monoxide (CO) - CO is a gas emitted directly from sources of equipment powered by fossil-fuels.

Parameter Estimates

AQI

_pm25 + _pm10 + _co

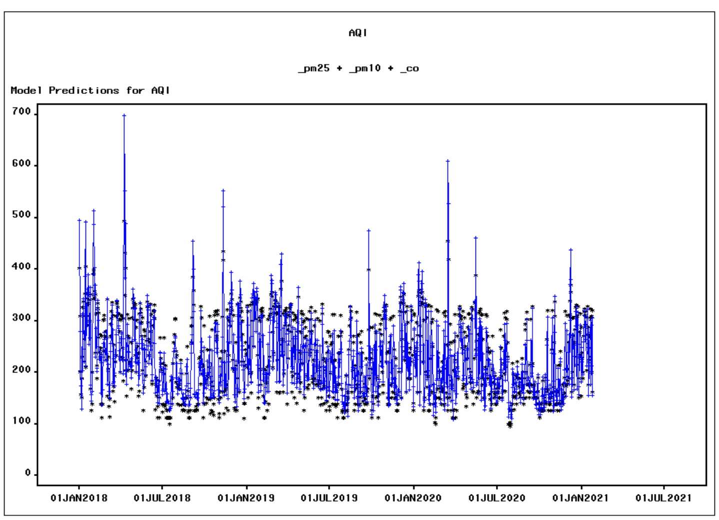
Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	28.43440	2.7498	10.3406	<.0001
_pm25	1.73453	0.0238	72.8620	<.0001
_pm10	0.20867	0.0614	3.4002	0.0010
_co	1.96234	0.4798	4.0901	<.0001
Model Variance (sigma squared)	687.43162	.	.	.

Fit Range: 01JAN2018 to 17OCT2020

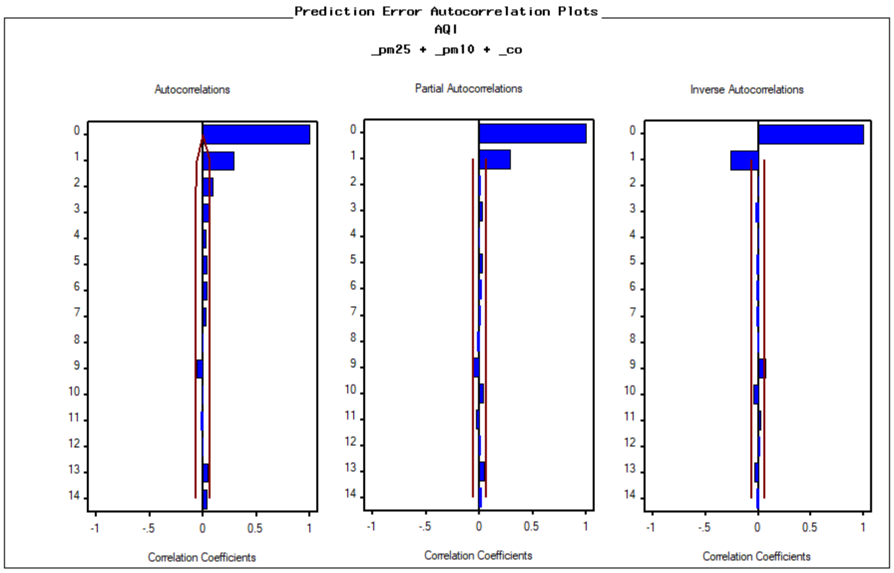
(Figure 21)

3.1 Regression Model and Analysis of Regression Residuals

With these three extra predictors, the prediction (fig. 22) now seems to be closer to the actual. All these three new predictors are significant (fig. 21). From the ACF (fig. 23), it shows a decay trend and coverage to 0 at around lag 4 quickly. Although the AQI itself shown in section 1 that it is not stationary, here now using the regression model, the remaining error is stationary. Therefore, the error model could still be built on this regression model. Since there exists a cut off at lag 1 in PACF, AR(1) might be suitable to fit in the error model.



(Figure 22)



(Figure 23)

3.2 Error Model using Regression Residuals

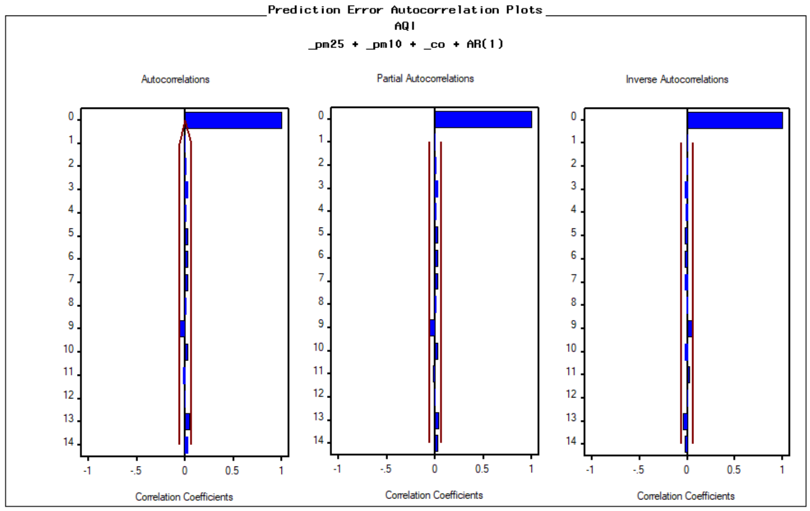
After fitting AR(1) in the error model, the plot (fig. 27) now seems to predict the actual well. From the ACF (fig. 24), no lag has value out of the bounds so the current error is now white noise. From the parameter estimation (fig. 26), all the parameters are significant. From the Statistics of Fit (fig. 24), MAPE is now 8.12867 which is smaller compared with models obtained in section 2.

Statistics of Fit
AQI
_pm25 + _pm10 + _co + AR(1)

Statistic of Fit	Value
Mean Square Error	458.58160
Root Mean Square Error	21.41452
Mean Absolute Percent Error	8.12867
Mean Absolute Error	17.07101
R-Square	0.913

Evaluation Range: 18OCT2020 to 25JAN2021

(Figure 24)



(Figure 25)

Parameter Estimates

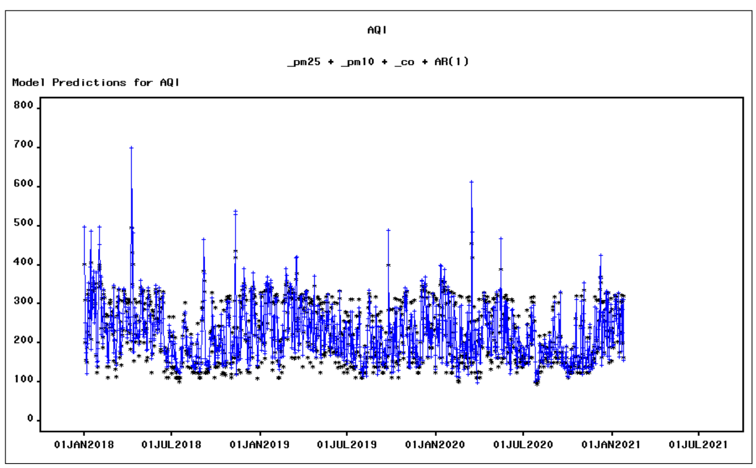
AQI

_pm25 + _pm10 + _co + AR(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	30.97209	3.3135	9.3474	<.0001
Autoregressive, Lag 1	0.29576	0.0301	9.8128	<.0001
_pm25	1.76198	0.0238	74.0357	<.0001
_pm10	0.14363	0.0670	2.1428	0.0347
_co	1.49956	0.5553	2.7006	0.0082
Model Variance (sigma squared)	631.11568	.	.	.

Fit Range: 01JAN2018 to 17OCT2020

(Figure 26)



(Figure 27)

3.3 Cross Correlation Analysis

In the cross-correlation, only lags with non-negative value are important to the prediction. As it has been shown in section 1, the dependent variable AQI is not stationary. From the autocorrelation of PM 2.5, PM 10 and CO (fig. 28-30), all of them show a slowing decaying trend and are not stationary. Therefore first differencing is needed for all three predictors and response before cross-correlation.

PM 2.5

The SAS System

The ARIMA Procedure

Name of Variable = _pm25	
Mean of Working Series	101.2961
Standard Deviation	39.45069
Number of Observations	1121
Embedded missing values in working series	3

Autocorrelations				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	1556.357	1.00000	| |********************|	0
1	819.416	0.52650	| .|*********** |	0.029867
2	372.034	0.23904	| .|***** |	0.037237
3	284.098	0.18254	| . |**** |	0.038582
4	240.309	0.15440	| . |*** |	0.039345
5	246.380	0.15831	| . |*** |	0.039882
6	223.015	0.14329	| . |*** |	0.040438
7	156.894	0.10081	| . |** |	0.040889
8	131.088	0.08423	| . |** |	0.041110
9	202.024	0.12981	| . |*** |	0.041264
10	224.134	0.14401	| . |*** |	0.041626

(Figure 28)

PM 10

The SAS System

The ARIMA Procedure

Name of Variable = _pm10	
Mean of Working Series	43.06798
Standard Deviation	18.3394
Number of Observations	1121
Embedded missing values in working series	3

Autocorrelations				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	336.333	1.00000	| |********************|	0
1	207.859	0.61802	| .|************ |	0.029867
2	113.196	0.33656	| . |******* |	0.039667
3	72.929212	0.21684	| . |**** |	0.042138
4	55.502823	0.16502	| . |*** |	0.043122
5	47.987886	0.14268	| . |*** |	0.043681
6	43.159074	0.12832	| . |*** |	0.044095
7	29.989398	0.08917	| . |** |	0.044427
8	22.627074	0.06728	| . |*. |	0.044586
9	43.149135	0.12829	| . |*** |	0.044677
10	50.082323	0.14891	| . |*** |	0.045004

(Figure 29)

CO

The SAS System

The ARIMA Procedure

<table>
<tr><th colspan="2">Name of Variable = _co</th></tr>
<tr><td>Mean of Working Series</td><td>5.694097</td></tr>
<tr><td>Standard Deviation</td><td>2.211812</td></tr>
<tr><td>Number of Observations</td><td>1121</td></tr>
<tr><td>Embedded missing values in working series</td><td>3</td></tr>
</table>

<table>
<tr><th colspan="5">Autocorrelations</th></tr>
<tr><th>Lag</th><th>Covariance</th><th>Correlation</th><th>-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1</th><th>Std Error</th></tr>
<tr><td>0</td><td>4.892112</td><td>1.00000</td><td>| |********************|</td><td>0</td></tr>
<tr><td>1</td><td>3.090723</td><td>0.63178</td><td>| . |************* |</td><td>0.029867</td></tr>
<tr><td>2</td><td>1.966953</td><td>0.40207</td><td>| . |******** |</td><td>0.040052</td></tr>
<tr><td>3</td><td>1.582921</td><td>0.32357</td><td>| . |****** |</td><td>0.043504</td></tr>
<tr><td>4</td><td>1.447448</td><td>0.29587</td><td>| . |****** |</td><td>0.045600</td></tr>
<tr><td>5</td><td>1.377295</td><td>0.28153</td><td>| . |****** |</td><td>0.047282</td></tr>
<tr><td>6</td><td>1.200539</td><td>0.24540</td><td>| . |***** |</td><td>0.048754</td></tr>
<tr><td>7</td><td>1.057781</td><td>0.21622</td><td>| . |**** |</td><td>0.049844</td></tr>
<tr><td>8</td><td>0.961463</td><td>0.19653</td><td>| . |**** |</td><td>0.050674</td></tr>
<tr><td>9</td><td>1.160460</td><td>0.23721</td><td>| . |***** |</td><td>0.051349</td></tr>
<tr><td>10</td><td>1.148277</td><td>0.23472</td><td>| . |***** |</td><td>0.052318</td></tr>
</table>

(Figure 30)

From the following results (fig. 31-33), lag 0, 1 and 2 are significant for PM2.5. Lag 1,2 and 3 are significant for PM 10. Lag 1 and3 are significant for CO.

PM 2.5

<table>
<tr><td>0</td><td>2559.604</td><td>0.90945</td><td>| . |****************** |</td></tr>
<tr><td>1</td><td>-531.682</td><td>-.18891</td><td>| ****| . |</td></tr>
<tr><td>2</td><td>-659.433</td><td>-.23430</td><td>| *****| . |</td></tr>
<tr><td>3</td><td>-70.944588</td><td>-.02521</td><td>| *| . |</td></tr>
<tr><td>4</td><td>-9.826610</td><td>-.00349</td><td>| . | . |</td></tr>
<tr><td>5</td><td>53.915940</td><td>0.01916</td><td>| . | . |</td></tr>
<tr><td>6</td><td>25.282891</td><td>0.00898</td><td>| . | . |</td></tr>
<tr><td>7</td><td>-116.643</td><td>-.04144</td><td>| *| . |</td></tr>
<tr><td>8</td><td>-196.770</td><td>-.06991</td><td>| *| . |</td></tr>
</table>

(Figure 31)

PM 10

<table>
<tr><td>0</td><td>-3.764452</td><td>-.00318</td><td>| . | . |</td></tr>
<tr><td>1</td><td>799.998</td><td>0.67499</td><td>| . |************* |</td></tr>
<tr><td>2</td><td>-215.168</td><td>-.18154</td><td>| ****| . |</td></tr>
<tr><td>3</td><td>-195.022</td><td>-.16455</td><td>| ***| . |</td></tr>
<tr><td>4</td><td>-10.291892</td><td>-.00868</td><td>| . | . |</td></tr>
<tr><td>5</td><td>-15.289594</td><td>-.01290</td><td>| . | . |</td></tr>
<tr><td>6</td><td>3.958818</td><td>0.00334</td><td>| . | . |</td></tr>
<tr><td>7</td><td>12.275538</td><td>0.01036</td><td>| . | . |</td></tr>
<tr><td>8</td><td>-34.854959</td><td>-.02941</td><td>| *| . |</td></tr>
</table>

(Figure 32)

CO

<table>
<tr><td>0</td><td>-15.267877</td><td>-.10906</td><td>| **|. |</td></tr>
<tr><td>1</td><td>84.578405</td><td>0.60413</td><td>| .|************ |</td></tr>
<tr><td>2</td><td>-1.792563</td><td>-.01280</td><td>| .|. |</td></tr>
<tr><td>3</td><td>-28.035546</td><td>-.20025</td><td>| ****|. |</td></tr>
<tr><td>4</td><td>-5.978140</td><td>-.04270</td><td>| *|. |</td></tr>
<tr><td>5</td><td>2.526999</td><td>0.01805</td><td>| .|. |</td></tr>
<tr><td>6</td><td>0.227011</td><td>0.00162</td><td>| .|. |</td></tr>
<tr><td>7</td><td>1.305454</td><td>0.00932</td><td>| .|. |</td></tr>
<tr><td>8</td><td>-0.513714</td><td>-.00367</td><td>| .|. |</td></tr>
</table>

(Figure 33)

After creating the correlated lag identified by cross-correlation, these lags are added into the first difference model as regressors. Considering the reality that the lag 0 of PM2.5, PM10 and CO will not be known until the prediction day for AQI, the following cross-correlation section will be divided into two parts: the first model that includes the significant lag 0 of predictors, the second model that does not includes the significant lag 0.

3.3.1 Estimation with Current Value of Predictors

From the parameter estimation (fig. 34), several of the parameters are not significant. Therefore, those non-significant regressors need to be removed.

Parameter Estimates

AQI

LAG1_PM25 + _pm25 + LAG2_PM25 + LAG1_PM10 + LAG2_PM10 + LAG3_PM10 + LAG1_CO + LAG3_CO + I(1)

Model Parameter	Estimate	Std. Error	T	Prob>¦T¦
Intercept	0.07893	0.9496	0.0831	0.9339
LAG1_PM25	-0.04592	0.0375	-1.2247	0.2239
_pm25	1.61990	0.0393	41.2498	<.0001
LAG2_PM25	0.05178	0.0389	1.3326	0.1860
LAG1_PM10	0.29559	0.0944	3.1298	0.0024
LAG2_PM10	0.01161	0.0870	0.1335	0.8941
LAG3_PM10	-0.10843	0.0949	-1.1431	0.2560
LAG1_CO	1.80858	0.7389	2.4478	0.0163
LAG3_CO	-2.09260	0.7447	-2.8100	0.0061
Model Variance (sigma squared)	899.02729	.	.	.

Fit Range: 04JAN2018 to 17OCT2020

(Figure 34)

Now the updated model has all significant regressors (fig. 35). From the ACF (fig. 36), it quickly decays to 0 and its PACF has a slowing decaying trend, This fits a MA model. Since it seems cut off at lag 2. An MA(2) or MA(1) would be applicable on the error model.

Parameter Estimates

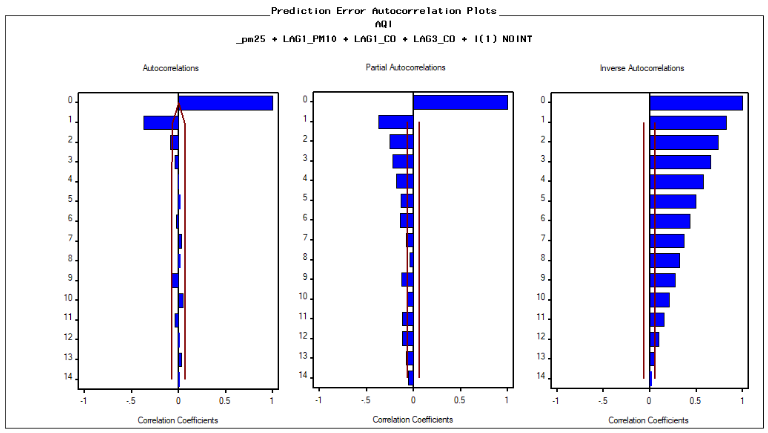
AQI

_pm25 + LAG1_PM10 + LAG1_CO + LAG3_CO + I(1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>¦T¦
_pm25	1.61029	0.0363	44.3768	<.0001
LAG1_PM10	0.30673	0.0901	3.4042	0.0010
LAG1_CO	1.99644	0.7300	2.7348	0.0074
LAG3_CO	-1.91145	0.5372	-3.5582	0.0006
Model Variance (sigma squared)	908.68061	.	.	.

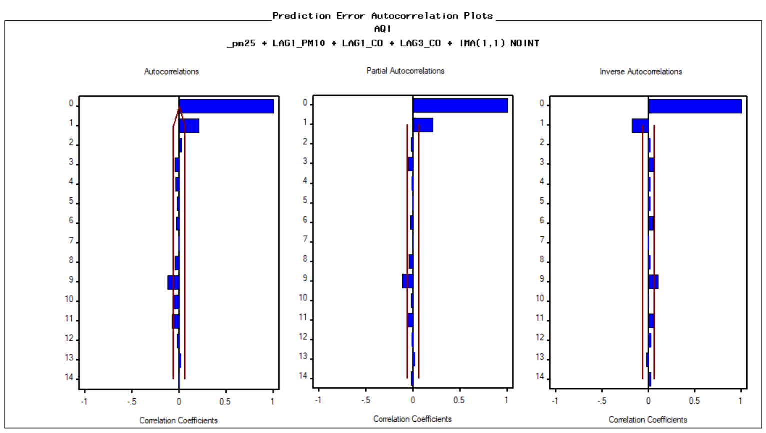
Fit Range: 04JAN2018 to 17OCT2020

(Figure 35)



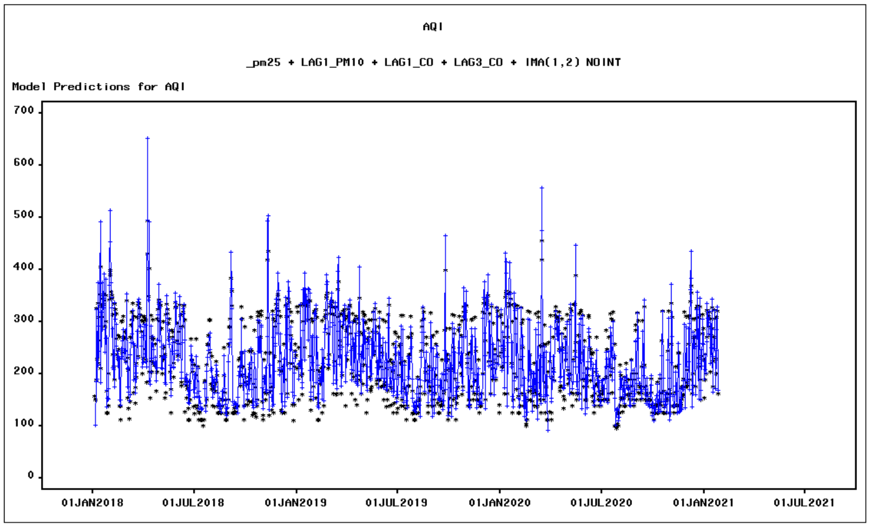
(Figure 36)

First, considering MA(1) for error model, the residuals after fitting MA(1) to the error are still not stationary. From the ACF (fig. 37), both ACF and PACF are significant at lag 1. Therefore, MA(1) might not be an optimal model for error.



(Figure 37)

Secondly, after fitting MA(2) into the error, the prediction now seems to fit the actual well (fig. 38). From the parameter estimation (fig. 39), Lag 3 for CO is slightly not significant, but besides that the rest parameters are strictly smaller than 5%. The error now seems to be a white noise since from the ACF (fig. 40) most of the lags are inside the two standard error bounds. In the validation set (fig. 41), this model has MAPE 9.00879 which is smaller compared with the models obtained in section 2 but slightly higher than section 3.1.



(Figure 38)

Parameter Estimates

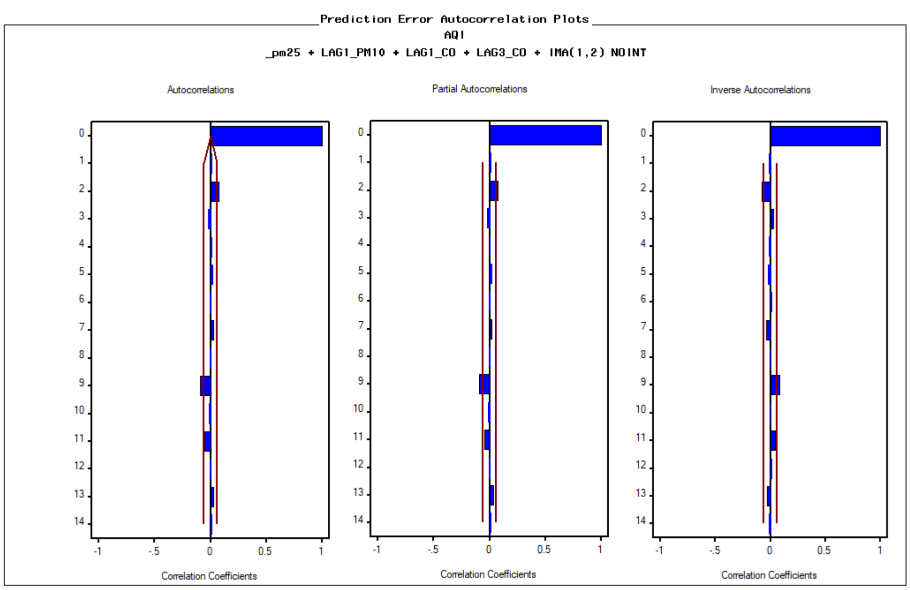
AQI

_pm25 + LAG1_PM10 + LAG1_CO + LAG3_CO + IMA(1,2) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.72519	0.0310	23.3602	<.0001
Moving Average, Lag 2	0.23623	0.0312	7.5652	<.0001
_pm25	1.56383	0.0349	44.8142	<.0001
LAG1_PM10	0.35535	0.0770	4.6161	<.0001
LAG1_CO	3.51862	0.6053	5.8128	<.0001
LAG3_CO	-0.86433	0.4435	-1.9490	0.0543
Model Variance (sigma squared)	605.40744	.	.	.

Fit Range: 04JAN2018 to 17OCT2020

(Figure 39)



(Figure 40)

Statistics of Fit

AQI

_pm25 + LAG1_PM10 + LAG1_CO + LAG3_CO + IMA(1,2) NOINT

Statistic of Fit	Value
Mean Square Error	584.50906
Root Mean Square Error	24.17662
Mean Absolute Percent Error	9.00879
Mean Absolute Error	19.18561
R-Square	0.889

Evaluation Range: 18OCT2020 to 25JAN2021

(Figure 41)

3.3.2 Estimation without Current Value of Predictors

Since lag 0 for PM 2.5 is the only significant lag 0 identified by the cross-correlation test, in this section it will not be considered as a regressor. The parameter estimation under this scenario is shown in (fig. 42). Intercept and the lag3 for CO have p-value larger than 0.05. Therefore, they should not be included in the model.

Parameter Estimates

AQI

LAG1_PM25 + LAG2_PM25 + LAG1_PM10 + LAG2_PM10 + LAG3_PM10 + LAG1_CO + LAG3_CO + I(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	-0.20867	1.5652	-0.1333	0.8942
LAG1_PM25	-0.48254	0.0593	-8.1350	<.0001
LAG2_PM25	-0.40358	0.0614	-6.5700	<.0001
LAG1_PM10	2.35600	0.1322	17.8191	<.0001
LAG2_PM10	0.59195	0.1415	4.1835	<.0001
LAG3_PM10	0.53948	0.1542	3.4991	0.0007
LAG1_CO	10.03179	1.1723	8.5572	<.0001
LAG3_CO	-0.88421	1.2255	-0.7215	0.4724
Model Variance (sigma squared)	2445	.	.	.

Fit Range: 04JAN2018 to 17OCT2020

(Figure 42)

After removing the insignificant terms, the remaining variables still have p-value smaller than 0.05 (fig. 43). The autocorrelation plot (fig. 44) seems that ACF is chopped off at lag 3 while the PACF shows a decaying trend. So MA(3) will be considered as a potential error model here.

Parameter Estimates

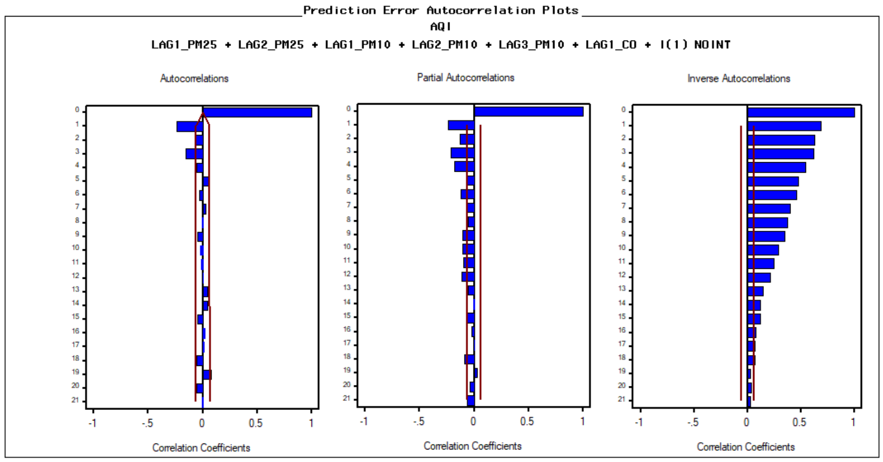
AQI

LAG1_PM25 + LAG2_PM25 + LAG1_PM10 + LAG2_PM10 + LAG3_PM10 + LAG1_CO + I(1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
LAG1_PM25	-0.48548	0.0591	-8.2169	<.0001
LAG2_PM25	-0.41630	0.0587	-7.0886	<.0001
LAG1_PM10	2.34620	0.1311	17.8955	<.0001
LAG2_PM10	0.58690	0.1411	4.1592	<.0001
LAG3_PM10	0.49211	0.1397	3.5224	0.0007
LAG1_CO	10.13988	1.1570	8.7641	<.0001
Model Variance (sigma squared)	2437	.	.	.

Fit Range: 04JAN2018 to 17OCT2020

(Figure 43)



(Figure 44)

The parameter estimation (fig. 45) after fitting MA(3) shows that all the regressors are still significant. Lag 2 for MA has a very large p-value but Lag 3 for MA is still significant. Furthermore, the residual ACF (fig. 46) now indicates the white noise series, Therefore, MA(3) is a suitable error model.

Parameter Estimates

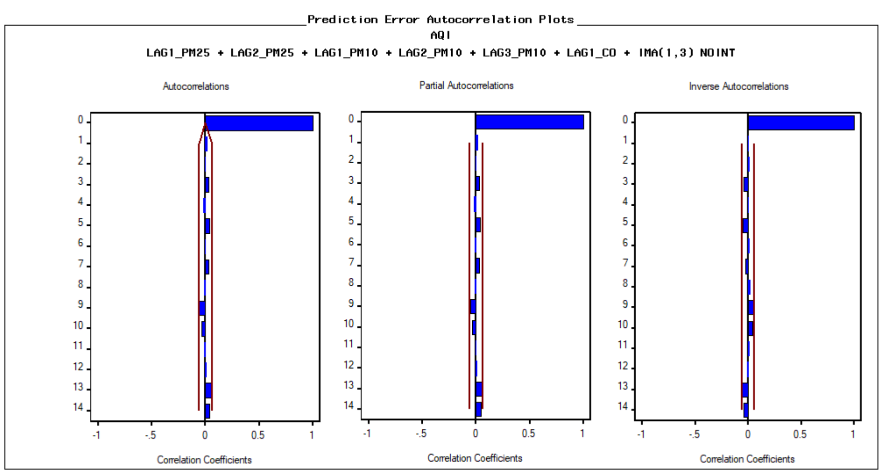
AQI

LAG1_PM25 + LAG2_PM25 + LAG1_PM10 + LAG2_PM10 + LAG3_PM10 + LAG1_CO + IMA(1,3) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.77071	0.0487	15.8172	<.0001
Moving Average, Lag 2	0.02753	0.0592	0.4651	0.6430
Moving Average, Lag 3	0.13871	0.0428	3.2385	0.0017
LAG1_PM25	0.25465	0.0877	2.9020	0.0046
LAG2_PM25	-0.27653	0.0806	-3.4297	0.0009
LAG1_PM10	2.29490	0.1218	18.8421	<.0001
LAG2_PM10	-0.58299	0.1777	-3.2802	0.0015
LAG3_PM10	0.45601	0.1620	2.8140	0.0060
LAG1_CO	11.83313	0.9968	11.8716	<.0001
Model Variance (sigma squared)	1774	.	.	.

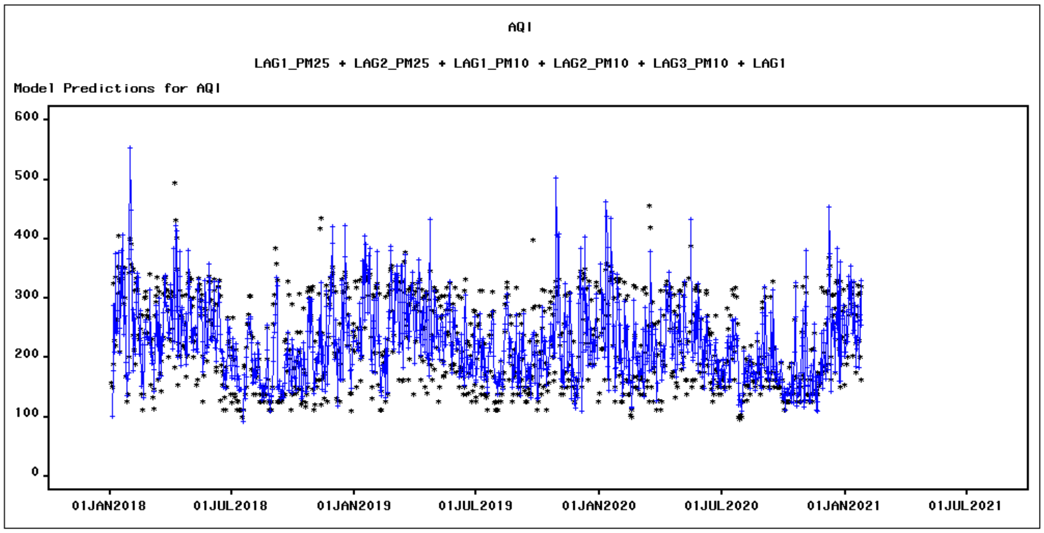
Fit Range: 04JAN2018 to 17OCT2020

(Figure 45)



(Figure 46)

From the prediction plot (fig. 47), it seems to fit the actual AQI quite well and correctly identified some of the extreme values. Based on the validation set, the current model now has MAPE 13.3460 (fig. 48).



(Figure 47)

Statistics of Fit

AQI

LAG1_PM25 + LAG2_PM25 + LAG1_PM10 + LAG2_PM10 + LAG3_PM10 + LAG1_CO + IMA(1,3) NOINT

Statistic of Fit	Value
Mean Square Error	1348.7
Root Mean Square Error	36.72476
Mean Absolute Percent Error	13.34596
Mean Absolute Error	29.02728
R-Square	0.744

Evaluation Range: 18OCT2020 to 25JAN2021

(Figure 48)

4. Conclusion

First, the fits of models are compared using the square root of model variance estimate. Based on Chart 49, two univariate models have the largest. The multivariate using only history records has a square root of model variance 42.12. While the other two multivariate models contain current or/and history records perform similarly and have the smallest.

Secondly, the comparison between models is based on the validation set using records from Oct 18th 2020 to Jan 25th 2021 with MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error).

Based on chart 50, the first two univariate models have large MAPE and MAE values compared with the following three multivariate models. Although the regression model (PM2.5 + PM10 + CO + Error AR(1)) has the lower MAPE which is about 8.1287, since it uses the current-day values of predictors, it would be not applicable in real life. The fourth model (PM2.5 + Lag1_PM10 + Lag1_CO + Lag3_CO + Error MA(2)) performs similarly with the regression model in both MAPE and MAE. Since it also includes current value of predictors during the estimation, this model would be selected considering actual application.

While for the last model (Lag1_PM2.5 + Lag2_PM2.5 + Lag1_PM10 + Lag2_PM10 +Lag1_CO + Error MA(3)) using only the history records for predictors, it has MAPE 13.3460 which is higher than other multivariate models that contain current records but still lower than the univariate models. So does its Therefore, Lag1_PM2.5 + Lag2_PM2.5 + Lag1_PM10 + Lag2_PM10 +Lag1_CO + Error MA(3) would be considered the best model for predicting AQI.

Models Fit Comparison	
Models	**Square Root of Model Variance Estimate**
Cyclical + Linear + Error AR(1)	70.74
ARIMA(0,1,2) Noint	65.15
PM2.5 + PM10 + CO + Error AR(1)	25.12
PM2.5 + Lag1_PM10 + Lag1_CO + Lag3_CO + Error MA(2)	24.61
Lag1_PM2.5 + Lag2_PM2.5 + Lag1_PM10 + Lag2_PM10 +Lag1_CO + Error MA(3)	42.12

(Chart 49)

Models Predictive Performance Comparison		
Models	**Mean Absolute Error**	**Mean Absolute Percent Error**
Cyclical + Linear + Error AR(1)	55.8994	26.9843
ARIMA(0,1,2) Noint	47.5923	21.3107
PM2.5 + PM10 + CO + Error AR(1)	17.0710	8.1287
PM2.5 + Lag1_PM10 + Lag1_CO + Lag3_CO + Error MA(2)	19.1856	9.0088
Lag1_PM2.5 + Lag2_PM2.5 + Lag1_PM10 + Lag2_PM10 +Lag1_CO + Error MA(3)	29.0273	13.34596

(Chart 50)