JiaHong Yu

1/1/2016

Research on Beijing ‘s PM 2.5 Prediction

Introduction:

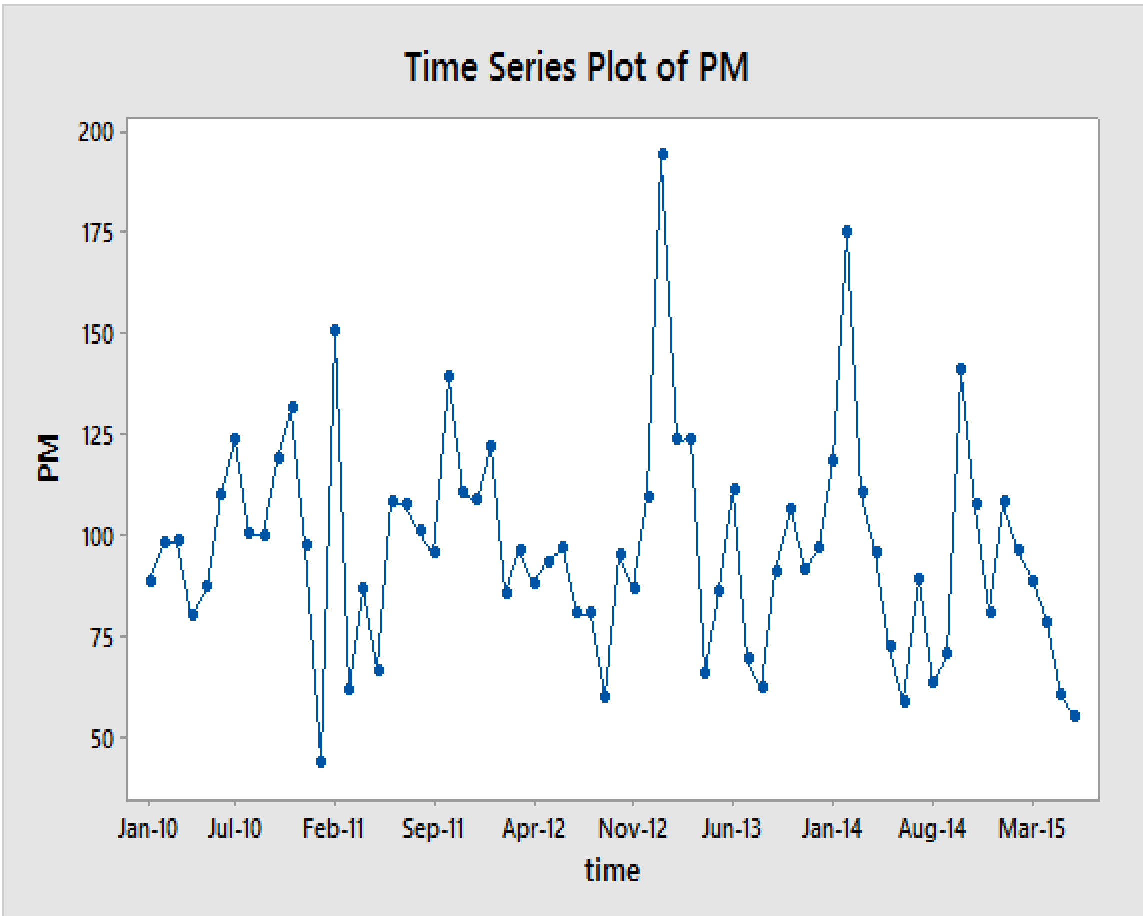
Beijing is the capital city of China which has severe air pollution. Being interested in how serious its pollution is and how the pollution level changes over time, I researched Beijing’s PM 2.5 hourly observations from Jan 1st, 2010 to Dec 31st, 2015 based on the data set ‘PM2.5 Data of Five Chinese Cities Data Set’ from UCI Machine Learning Repository. My research offers a solid overview of the air pollution in Beijing, which can help the Beijing government design its pollution control policies in 2016.

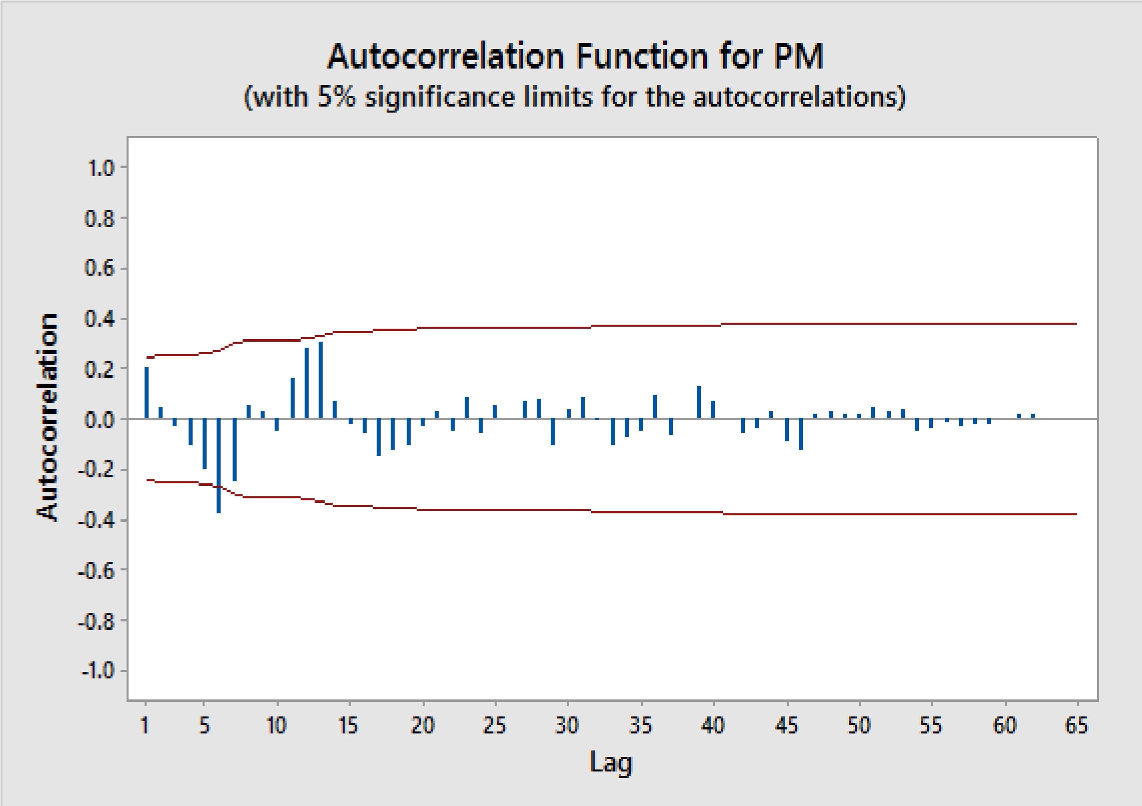
The original data set had many missing values named NA. To model time series I needed evenly spaced data. Therefore, a function called ‘na.ma’ in RStudio was used to fix these NAs. This function could replace the missing values based on a weighted exponential moving average. However, at some points in time, the wind direction information was missed and could not be fixed by ‘na.ma’ function because they were categorical values. Therefore, I manually fixed these 'NA's based on the wind directions around these points in time. In the end, all the NAs in this data set were fixed.

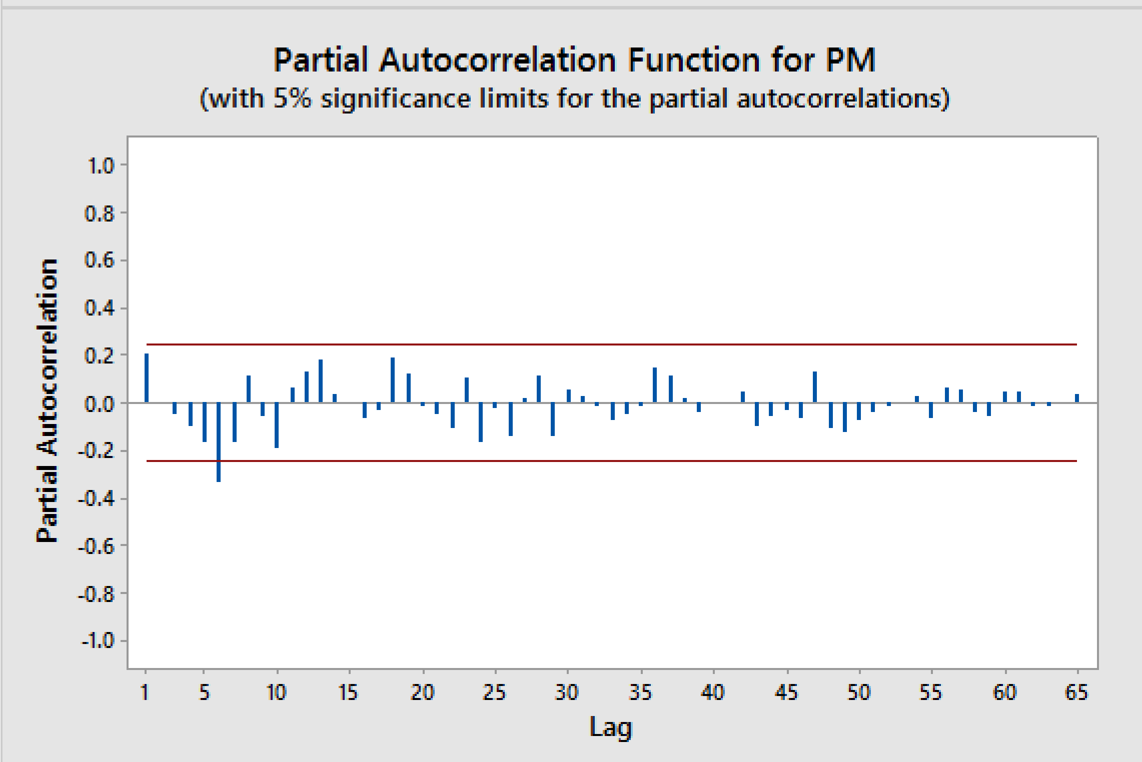
I made two subsets of the raw data—daily observations, and monthly observations. The daily observations were made by subsetting the observations at noon of each day; meanwhile, the monthly observations were obtained by calculating the average PM 2.5 of each month. For each subset, the last few records were held as the out-of-sample observations.

Monthly Data Analysis:

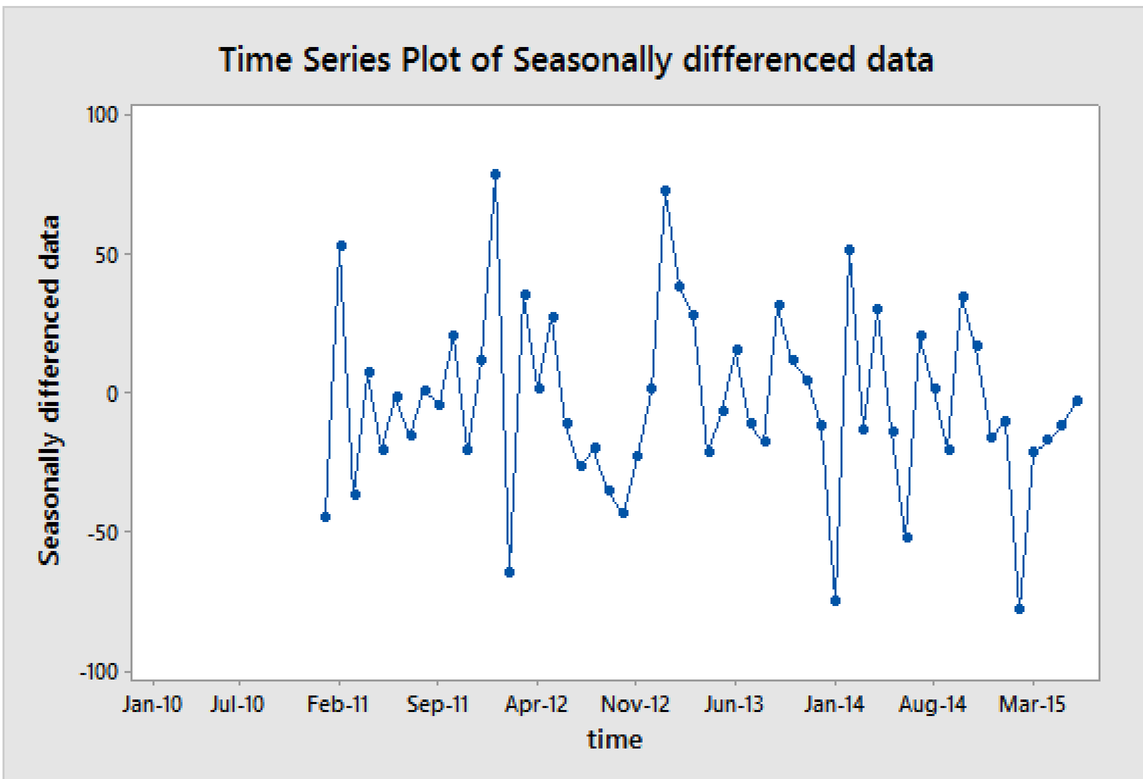
The monthly observations were first analyzed in Minitab 7. I loaded the monthly subset and drew its time series, ACF, and PACF plots. The plots are displayed below:



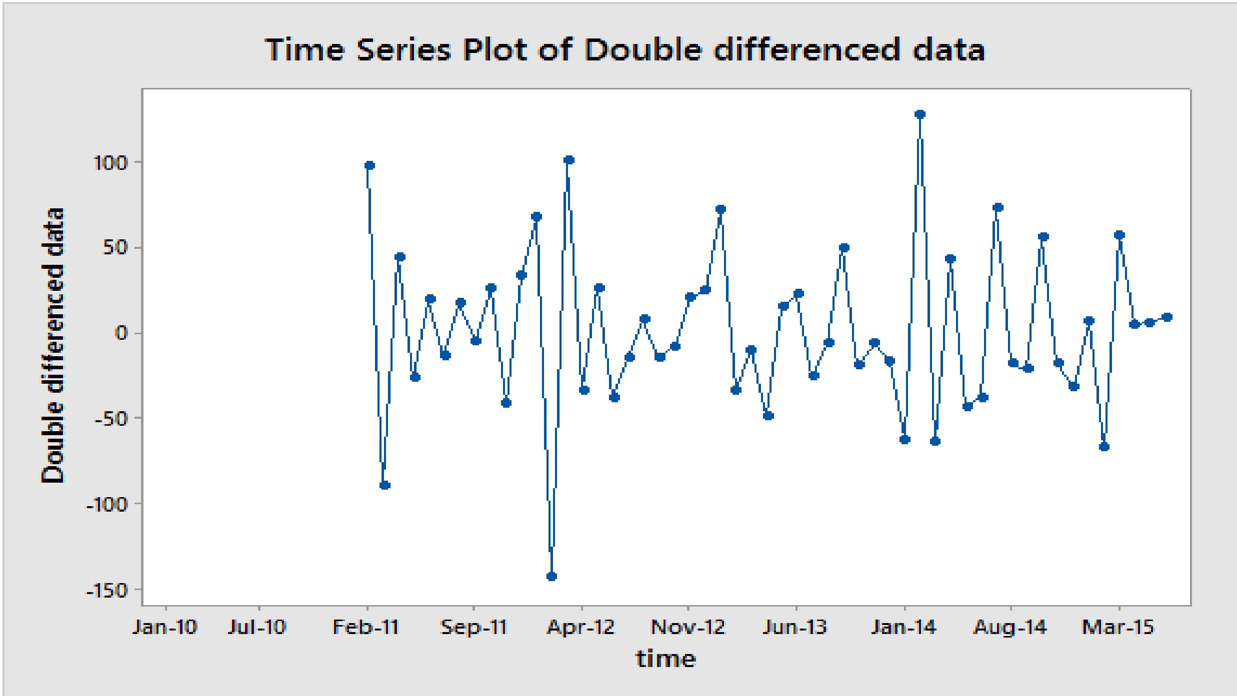

Autocorrelation Function for PM
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 5 10 15 20 25 30 35 40 45 50 55 60 65
Lag

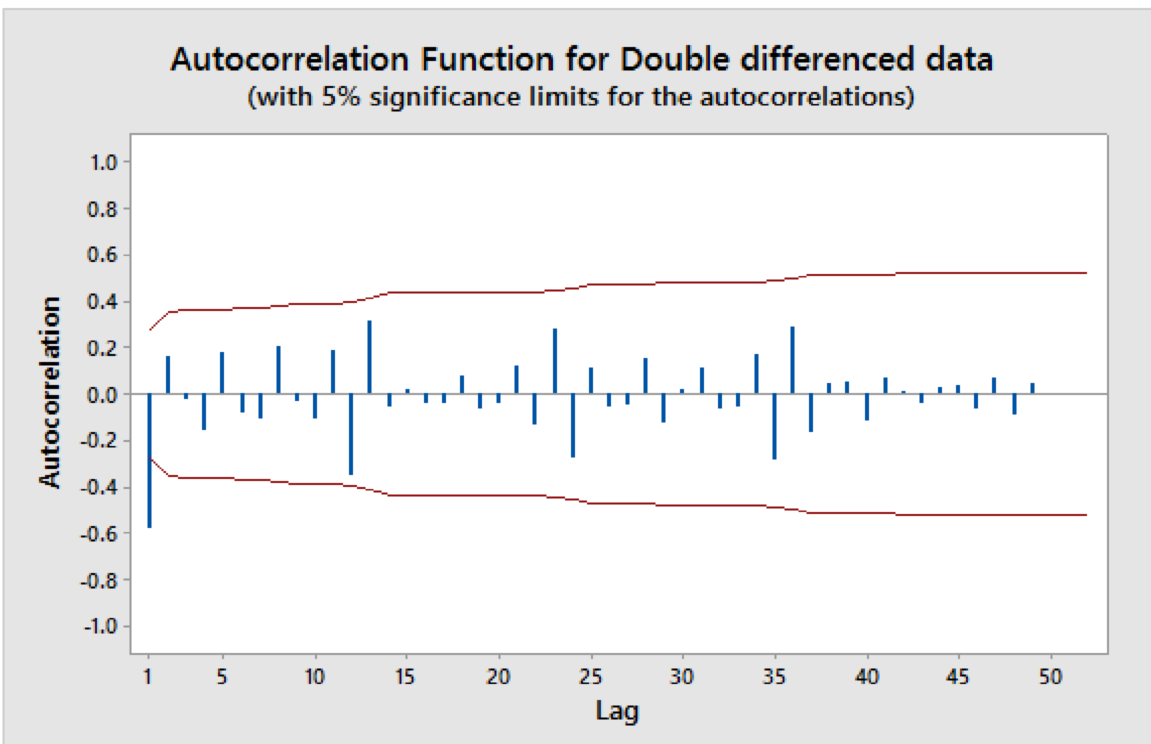

Partial Autocorrelation Function for PM
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 5 10 15 20 25 30 35 40 45 50 55 60 65
Lag

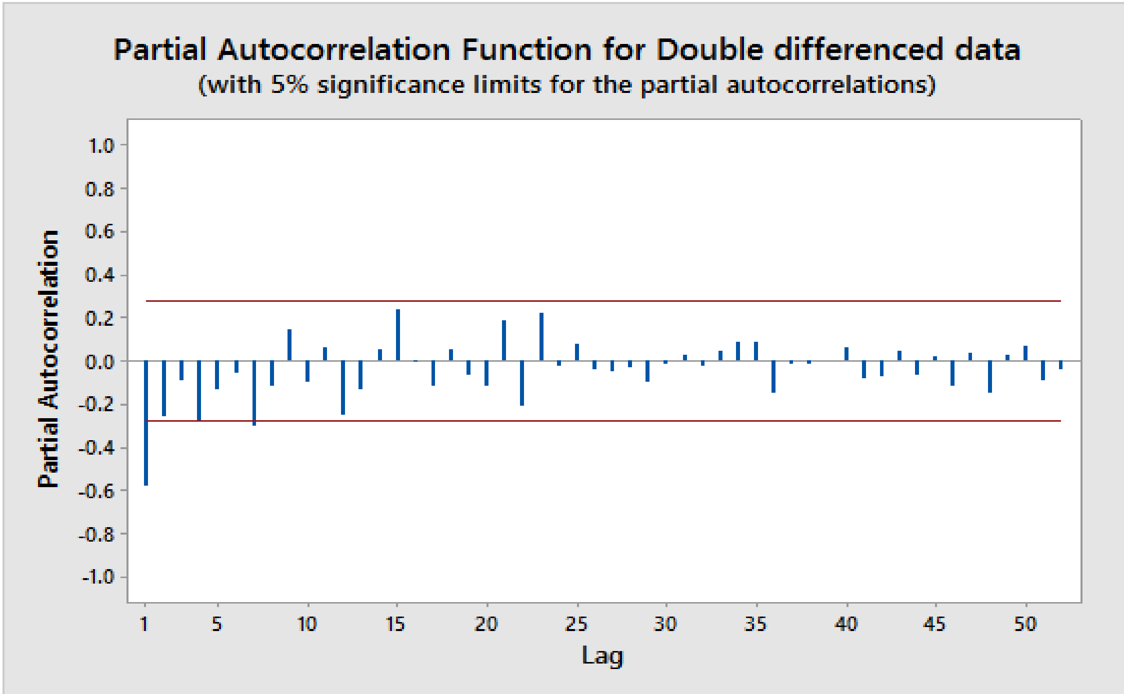
As we could see, there are apparent seasonal patterns in all above graphs: about every 12 months, Beijing's PM 2.5 is strongly positive-correlated and about every six months, Beijing's PM 2.5 is strongly negative-correlated. This phenomenon is because Beijing's air pollution usually reaches its highest level in winter and reaches its lowest level in summer. To confirm the seasonal pattern, I made a time series plot for the seasonally differenced data:



As depicted above, this plot still does not have a stationary mean. Therefore, I applied the first difference to the seasonally differenced data:


Time Series Plot of Double differenced data
Double differenced data
100
50
0
-50
-100
-150
Jan-10
Jul-10
Feb-11
Sep-11
Apr-12
Nov-12
Jun-13
Jan-14
Aug-14
Mar-15
time


Autocorrelation Function for Double differenced data
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
5
10
15
20
25
30
35
40
45
50
Lag



Now the above plot is stationary. There are also significant spikes at lag 1 in ACF and PACF plots while both graphs are sinusoidal. Though both plots have shown some seasonality as their autocorrelations exponentially decay at lag 12, lag 24, and lag 36, none of them show any significant spike at lag 12. Therefore, the best candidate model for the monthly PM 2.5 observations of Beijing is around Seasonal ARIMA (1,1,0) $(0,1,0)_{12}$ or Seasonal ARIMA (0,1,1) $(0,1,0)_{12}$.

The Minitab worksheet was saved as a CSV file and opened in JMP. Because the monthly subset was estimated by the mean value of each month's PM 2.5 observations, I had to drop the other variables that may have affected Beijing's PM 2.5, like wind direction, as they were vectors and could not be averaged. Only a time series ARIMA could be made for the monthly PM 2.5 observations.

After multiple attempts, the two best ARIMA models from JMP were Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept and Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept. Their residuals' ACF and PACF plots are shown below:

a. Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
1	0.0058	0.0019	0.9656	1	0.0058
2	0.1025	0.6029	0.7397	2	0.1025
3	-0.0775	0.9529	0.8126	3	-0.0794
4	-0.1527	2.3402	0.6735	4	-0.1645
5	-0.0641	2.5899	0.7629	5	-0.0480
6	-0.1938	4.9190	0.5542	6	-0.1728
7	-0.1348	6.0710	0.5315	7	-0.1623
8	0.1094	6.8460	0.5533	8	0.1154
9	0.0126	6.8564	0.6521	9	-0.0028
10	-0.0603	7.1027	0.7157	10	-0.1862
11	0.0993	7.7872	0.7322	11	0.0626
12	0.0276	7.8414	0.7974	12	0.0491
13	0.2463	12.2633	0.5062	13	0.1832
14	0.0921	12.8967	0.5347	14	0.1205
15	0.0113	12.9064	0.6095	15	0.0351
16	-0.0741	13.3393	0.6478	16	-0.1100
17	-0.1129	14.3705	0.6407	17	-0.0501
18	-0.0595	14.6655	0.6848	18	0.0824
19	-0.0969	15.4703	0.6923	19	-0.0103
20	-0.0502	15.6928	0.7355	20	-0.0272
21	0.0865	16.3750	0.7482	21	0.0697
22	-0.0104	16.3852	0.7963	22	-0.1011
23	0.2207	21.1163	0.5740	23	0.2053
24	-0.1380	23.0300	0.5180	24	-0.1433
25	0.0426	23.2192	0.5648	25	-0.0210

b. Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept.

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
1	-0.1040	0.6067	0.4360	1	-0.1040
2	-0.1816	2.4926	0.2876	2	-0.1946
3	-0.0432	2.6015	0.4572	3	-0.0904
4	-0.1094	3.3134	0.5068	4	-0.1720
5	0.0488	3.4582	0.6297	5	-0.0192
6	-0.1789	5.4427	0.4884	6	-0.2584
7	-0.1069	6.1666	0.5204	7	-0.2189
8	0.2264	9.4871	0.3029	8	0.0581
9	0.0147	9.5015	0.3923	9	-0.0571
10	-0.1054	10.2541	0.4185	10	-0.1695
11	-0.0064	10.2569	0.5074	11	-0.0952
12	0.0482	10.4224	0.5790	12	-0.0321
13	0.2324	14.3595	0.3490	13	0.1667
14	0.0546	14.5827	0.4073	14	0.1805
15	-0.0505	14.7785	0.4675	15	0.1777
16	-0.0953	15.4944	0.4888	16	-0.0101
17	-0.1019	16.3358	0.5002	17	-0.0363
18	-0.0466	16.5164	0.5566	18	0.0029
19	-0.0925	17.2493	0.5730	19	-0.0464
20	-0.0066	17.2532	0.6365	20	-0.0401
21	0.1193	18.5488	0.6141	21	-0.0137
22	0.0574	18.8586	0.6540	22	-0.0557
23	0.1961	22.5938	0.4847	23	0.2085
24	-0.1416	24.6105	0.4272	24	-0.0530
25	-0.0014	24.6107	0.4844	25	0.0664

Both the ACF and PACF plots are very clean because no autocorrelation coefficient exceeds their threshold. This phenomenon means, there is no autocorrelation in residuals under both models.

a. Seasonal ARIMA $(0,1,1)(0,1,1)_{12}$ No Intercept

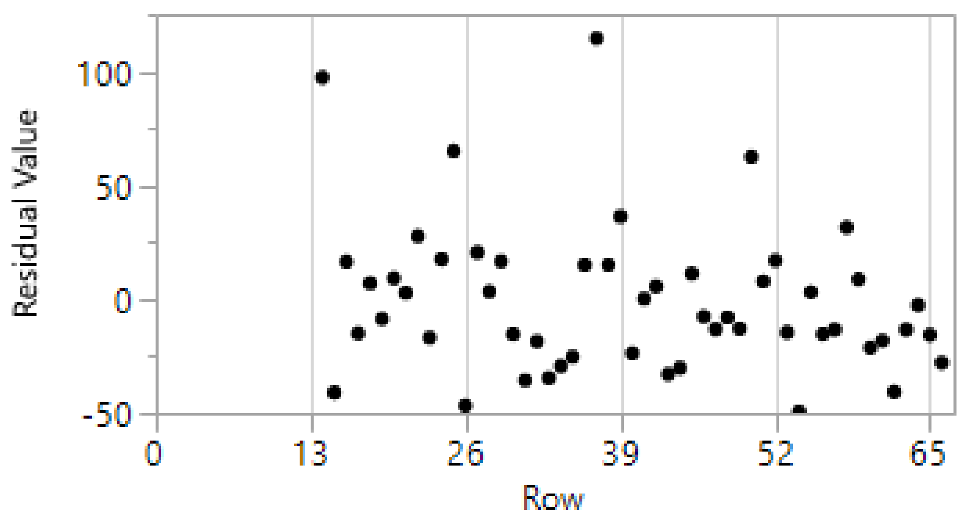
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
MA1,1	1	1	0.99992185	0.1906786	5.24	<.0001*
MA2,12	2	12	0.60379246	0.2003832	3.01	0.0040*

b. Seasonal ARIMA $(1,1,0)(0,1,1)_{12}$ No Intercept

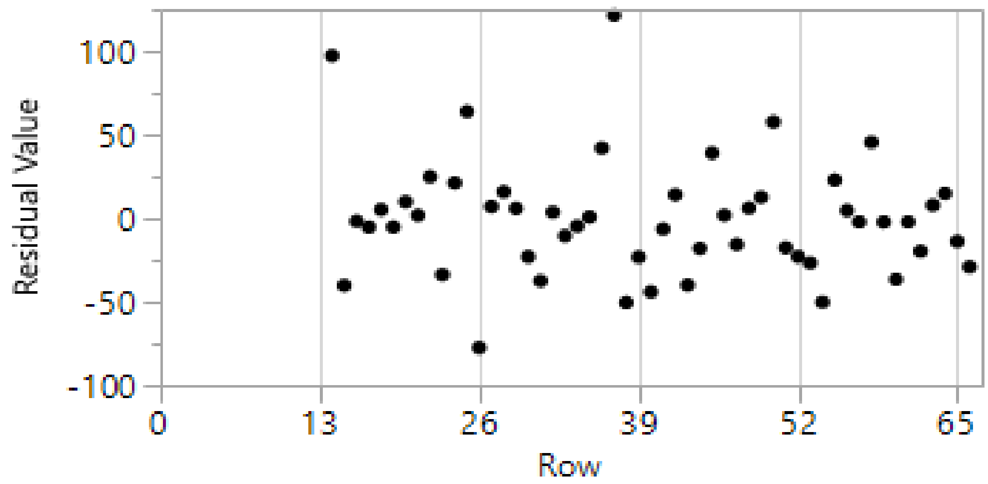
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	-0.5079423	0.1163427	-4.37	<.0001*
MA2,12	2	12	0.9999065	0.4981053	2.01	0.0500

Both models are approximately significant as the p-values of their ARIMA coefficients are smaller or equal to 0.05.

a. Seasonal ARIMA $(0,1,1)(0,1,1)_{12}$ No Intercept



b. Seasonal ARIMA $(1,1,0)(0,1,1)_{12}$ No Intercept



The residual plot is not very perfect as there are some outliers on the top of both models' residual plots. This phenomenon is most likely because the Chinese Spring Festival at the beginning of each year always involves a lot of fireworks.

All above analysis proved that Seasonal ARIMA $(0,1,1)(0,1,1)_{12}$ No Intercept and Seasonal ARIMA $(1,1,0)(0,1,1)_{12}$ No Intercept fit the monthly PM 2.5 observations of Beijing very well. However, I needed to compare my model's forecasts with the real out-of-sample observations in the last half year of 2015(6 months).

Seasonal ARIMA$(0,1,1)(0,1,1)_{12}$:

Month	Real PM Observations	Simple Forecast PM	Simple Residuals	Squared Error	MSE
67	55.03	82.03	-27.00	729.00	1704.62
68	44.64	67.19	-22.55	508.50	
69	47.10	73.33	-26.23	688.01	
70	72.09	117.55	-45.47	2067.52	
71	124.82	96.85	27.97	782.32	
72	161.97	88.13	73.84	5452.35	

Seasonal ARIMA(0,1,1)(1,1,0)12:

Month	Real PM Observations	Simple Forecast PM	Simple Residuals	Squared Error	MSE
67	55.03	62.5119	-7.4819	55.98	2138.63
68	44.64	44.5393	0.1007	0.01	
69	47.10	49.2214	-2.1214	4.50	
70	72.09	84.6882	-12.5982	158.72	
71	124.82	70.6568	54.1632	2933.65	
72	161.97	63.5885	98.3815	9678.91	

We could see, Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept's MSE(1704.62) is much lower than Seasonal ARIMA (1,1,0)(0,1,1)12 No Intercept's(2138.63). Therefore, Seasonal ARIMA (0,1,1)(0,1,1)12 No Intercept is the best available model to the forecast monthly average PM 2.5 of Beijing.

The explicit model equations of Seasonal ARIMA (0,1,1)(0,1,1)12 are given below:

$$(1-B)(1-B^{12})\, y_t = (1-0.9999\mathrm{B})(1-0.6038B^{12})\, e_t$$

$$y_t = y_{t-1} + y_{t-12} - y_{t-13} + e_t - 0.9999\,e_{t-1} - 0.6038e_{t-12} + 0.6037\,e_{t-13}$$

Therefore, the forecast equations for the next three periods are as follows:

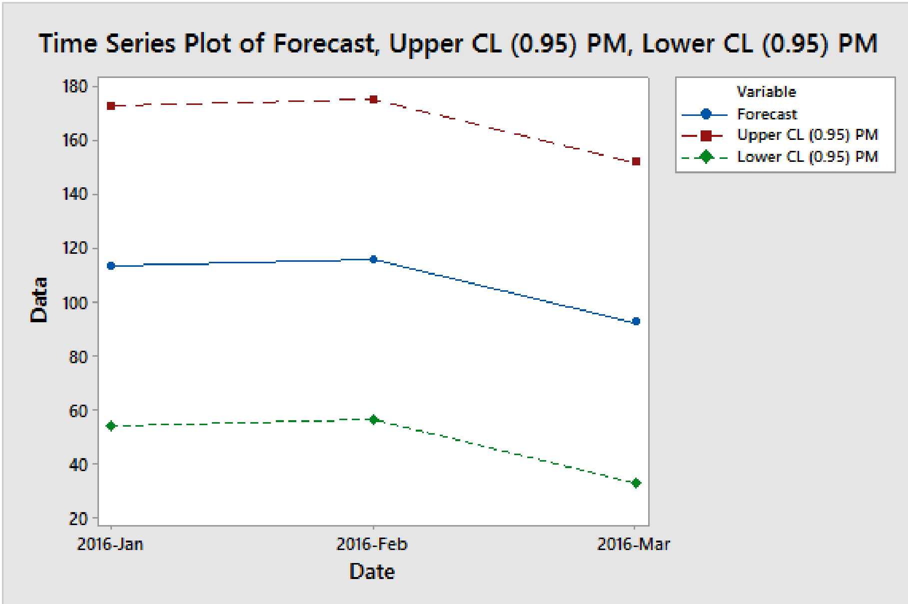
$$\hat{y}_{t+1} = y_t + y_{t-11} - y_{t-12} + 0 - 0.9999\,e_t - 0.6038e_{t-11} + 0.6037e_{t-12}$$

$$\hat{y}_{t+2} = \hat{y}_{t+1} + y_{t-10} - y_{t-11} + 0 - 0 - 0.6038e_{t-10} + 0.6037e_{t-11}$$

$$\hat{y}_{t+3} = \hat{y}_{t+2} + y_{t-9} - y_{t-10} + 0 - 0 - 0.6038\,e_{t-9} + 0.6037\,e_{t-10}$$

Based on the chosen model, I combined the in-sample and out-of-sample monthly observations and forecasted the average PM 2.5 value of the next year's first three months (2016-Jan, 2016-Feb, 2016-Mar) with a 95% confidence interval:

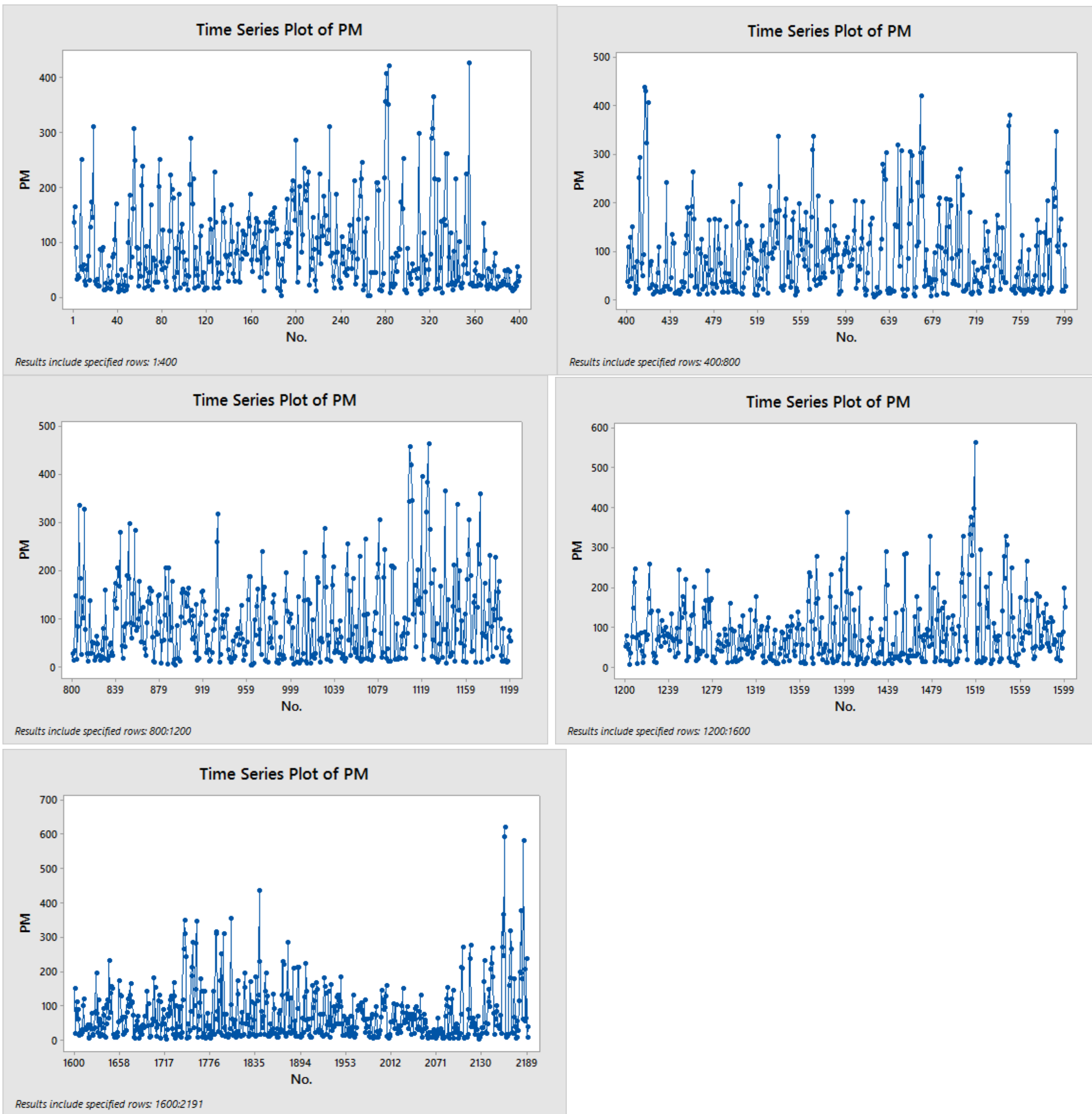
Data	Period	Forecast	Upper CL (0.95) PM	Lower CL (0.95) PM
2016-Jan	73	113.147	172.606	53.6885
2016-Feb	74	115.546	175.005	56.0875
2016-Mar	75	92.186	151.645	32.7279



From the time plot, Beijing's monthly PM 2.5 value would decrease in the next 3 months.

Daily Data Analysis:

I then started to study the forecast model for the daily PM 2.5 observations of Beijing. First, Minitab 7 was used to make the time series plot:



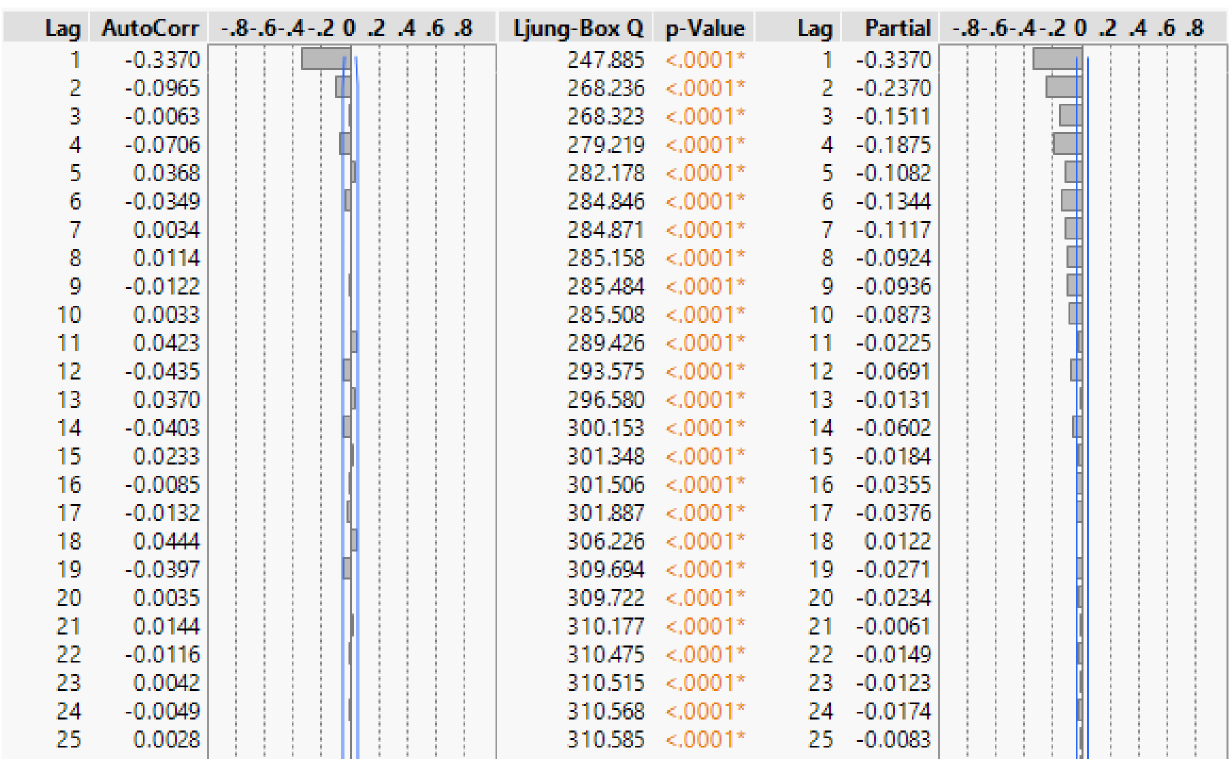
From the time series plot, there is no clear trend or seasonality of PM2.5 observation data. Thus, I used R to analyze this data set.

ARIMA models only include information from the past observations of a series. To make predictions for the future PM 2.5, Some other useful information like Dew Point (Celsius Degree), Temperature (Celsius Degree), Humidity (%), Pressure (hPa), Cumulative wind speed (m/s), and hourly precipitation (mm) are needed in the data set.

According to the KPSS test in R, PM 2.5 and Cumulative wind speed are not stationary. The p-values of each variable are given below:

PM	DEWP	HUMI	PRES	TEMP	Iws	precipitation
0.04155	0.1	0.1	0.1	0.1	0.01093	0.1

Therefore, I needed to difference the data first and then applied the ordinary regression to the differenced variables in JMP. The ACF and PACF plots of the regression residuals are given below:

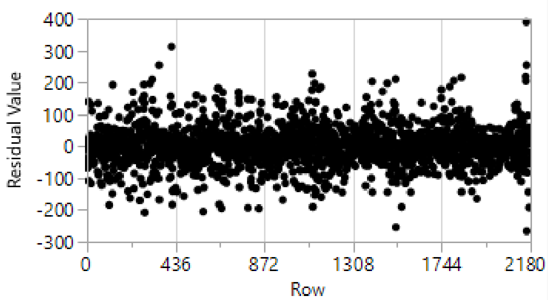


Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	-0.3370		247.885	<.0001*	1	-0.3370	
2	-0.0965		268.236	<.0001*	2	-0.2370	
3	-0.0063		268.323	<.0001*	3	-0.1511	
4	-0.0706		279.219	<.0001*	4	-0.1875	
5	0.0368		282.178	<.0001*	5	-0.1082	
6	-0.0349		284.846	<.0001*	6	-0.1344	
7	0.0034		284.871	<.0001*	7	-0.1117	
8	0.0114		285.158	<.0001*	8	-0.0924	
9	-0.0122		285.484	<.0001*	9	-0.0936	
10	0.0033		285.508	<.0001*	10	-0.0873	
11	0.0423		289.426	<.0001*	11	-0.0225	
12	-0.0435		293.575	<.0001*	12	-0.0691	
13	0.0370		296.580	<.0001*	13	-0.0131	
14	-0.0403		300.153	<.0001*	14	-0.0602	
15	0.0233		301.348	<.0001*	15	-0.0184	
16	-0.0085		301.506	<.0001*	16	-0.0355	
17	-0.0132		301.887	<.0001*	17	-0.0376	
18	0.0444		306.226	<.0001*	18	0.0122	
19	-0.0397		309.694	<.0001*	19	-0.0271	
20	0.0035		309.722	<.0001*	20	-0.0234	
21	0.0144		310.177	<.0001*	21	-0.0061	
22	-0.0116		310.475	<.0001*	22	-0.0149	
23	0.0042		310.515	<.0001*	23	-0.0123	
24	-0.0049		310.568	<.0001*	24	-0.0174	
25	0.0028		310.585	<.0001*	25	-0.0083	

Obviously, the residuals are strongly autocorrelated. Therefore, it was necessary to fit autocorrelated regression noise instead with ARMA models. Based on the ACF and PACF plots of the ordinary regression residuals, multiple different models were tried in JMP and the two best candidates were: ARMA(1, 1) and ARMA (2, 2):

ARMA(1, 1)

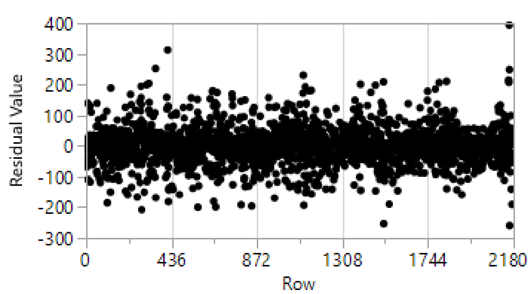
Variable	Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
DEWP	Num0,0	0	0	4.995141	0.450213	11.10	<.0001*
HUMI	Num0,0	0	0	0.992955	0.127525	7.79	<.0001*
PRES	Num0,0	0	0	-1.262543	0.294833	-4.28	<.0001*
Iws	Num0,0	0	0	-0.078908	0.027840	-2.83	0.0046*
precipitation	Num0,0	0	0	-8.804075	2.465855	-3.57	0.0004*
PM	AR1,1	1	1	0.317011	0.024283	13.06	<.0001*
PM	MA1,1	1	1	0.889294	0.011075	80.30	<.0001*



Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	-0.0003		0.0002	0.9900	1	-0.0003	
2	0.0045		0.0447	0.9779	2	0.0045	
3	0.0072		0.1572	0.9842	3	0.0072	
4	-0.0582		7.5528	0.1094	4	-0.0582	
5	0.0089		7.7241	0.1721	5	0.0088	
6	-0.0384		10.9454	0.0901	6	-0.0380	
7	-0.0068		11.0469	0.1366	7	-0.0060	
8	0.0060		11.1268	0.1946	8	0.0028	
9	-0.0023		11.1380	0.2664	9	-0.0007	
10	0.0179		11.8374	0.2961	10	0.0135	
11	0.0432		15.9308	0.1437	11	0.0433	
12	-0.0170		16.5637	0.1668	12	-0.0180	
13	0.0323		18.8559	0.1276	13	0.0314	
14	-0.0222		19.9418	0.1320	14	-0.0209	
15	0.0234		21.1467	0.1322	15	0.0284	
16	0.0034		21.1723	0.1720	16	0.0013	
17	0.0023		21.1840	0.2182	17	0.0101	
18	0.0423		25.1280	0.1214	18	0.0382	
19	-0.0166		25.7373	0.1377	19	-0.0116	
20	0.0157		26.2834	0.1567	20	0.0142	
21	0.0256		27.7240	0.1482	21	0.0261	
22	0.0079		27.8599	0.1804	22	0.0117	
23	0.0206		28.7962	0.1872	23	0.0192	
24	0.0168		29.4215	0.2047	24	0.0192	
25	0.0279		31.1445	0.1843	25	0.0323	

ARMA(2, 2)

Variable	Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
DEWP	Num0,0	0	0	4.999452	0.417646	11.97	<.0001*
HUMI	Num0,0	0	0	0.977542	0.121733	8.03	<.0001*
PRES	Num0,0	0	0	-1.250985	0.292626	-4.28	<.0001*
Iws	Num0,0	0	0	-0.074230	0.027337	-2.72	0.0067*
precipitation	Num0,0	0	0	-8.749210	2.450251	-3.57	0.0004*
PM	AR1,1	1	1	-0.657716	0.001222	-538.4	<.0001*
PM	AR1,2	1	2	0.297343	0.000577	515.58	<.0001*
PM	MA1,1	1	1	-0.092339	0.000180	-514.3	<.0001*
PM	MA1,2	1	2	0.871457	0.001601	544.35	<.0001*

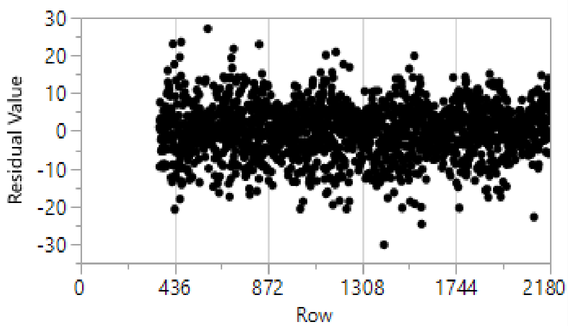


Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	-0.0034		0.0259	0.8722	1	-0.0034	
2	0.0174		0.6878	0.7090	2	0.0174	
3	-0.0045		0.7319	0.8657	3	-0.0044	
4	-0.0478		5.7361	0.2197	4	-0.0482	
5	-0.0027		5.7524	0.3311	5	-0.0029	
6	-0.0288		7.5693	0.2714	6	-0.0272	
7	-0.0170		8.2053	0.3148	7	-0.0176	
8	0.0147		8.6811	0.3699	8	0.0133	
9	-0.0112		8.9573	0.4412	9	-0.0111	
10	0.0259		10.4277	0.4038	10	0.0226	
11	0.0347		13.0758	0.2884	11	0.0338	
12	-0.0088		13.2469	0.3514	12	-0.0091	
13	0.0237		14.4829	0.3407	13	0.0209	
14	-0.0142		14.9249	0.3833	14	-0.0107	
15	0.0153		15.4361	0.4205	15	0.0176	
16	0.0104		15.6745	0.4759	16	0.0114	
17	-0.0046		15.7201	0.5438	17	-0.0002	
18	0.0483		20.8549	0.2868	18	0.0473	
19	-0.0229		22.0065	0.2839	19	-0.0205	
20	0.0213		23.0070	0.2885	20	0.0214	
21	0.0193		23.8272	0.3015	21	0.0188	
22	0.0138		24.2467	0.3344	22	0.0186	
23	0.0150		24.7418	0.3637	23	0.0124	
24	0.0221		25.8183	0.3624	24	0.0255	
25	0.0225		26.9398	0.3589	25	0.0259	

As seen from the results, all residuals are clean and all model parameters are significant. To choose a better model, I forecasted the out-of-sample data (12/22/2015 - 12/31/2015) and calculated the out-of-sample MSE. An ARIMA model had to be fitted for each predictor in JMP:

Dew Point: Seasonal ARIMA(2,0,0)(0,1,1)365 without intercept

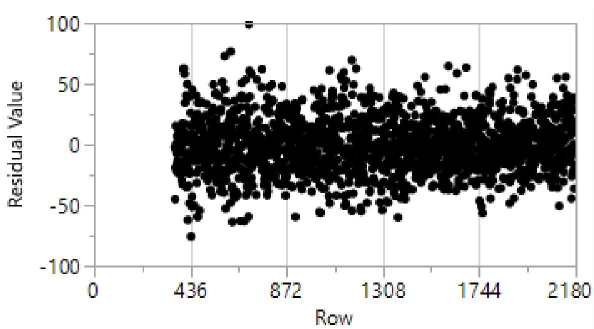
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	0.5188903	0.0234485	22.13	<.0001*
AR1,2	1	2	-0.0521165	0.0234835	-2.22	0.0266*
MA2,365	2	365	0.9989377	0.1044136	9.57	<.0001*



Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	0.0005		0.0004	0.9843	1	0.0005	
2	-0.0075		0.1025	0.9500	2	-0.0075	
3	-0.0071		0.1936	0.9786	3	-0.0071	
4	-0.0015		0.1978	0.9954	4	-0.0016	
5	-0.0087		0.3349	0.9969	5	-0.0088	
6	0.0047		0.3750	0.9990	6	0.0046	
7	0.0332		2.3861	0.9354	7	0.0330	
8	0.0449		6.0665	0.6398	8	0.0449	
9	-0.0063		6.1386	0.7260	9	-0.0057	
10	0.0081		6.2571	0.7932	10	0.0091	
11	0.0088		6.4000	0.8454	11	0.0096	
12	0.0723		15.9621	0.1930	12	0.0733	
13	0.0276		17.3598	0.1834	13	0.0288	
14	-0.0117		17.6118	0.2250	14	-0.0120	
15	0.0462		21.5209	0.1210	15	0.0453	
16	0.0109		21.7387	0.1518	16	0.0104	
17	0.0355		24.0555	0.1179	17	0.0381	
18	0.0093		24.2148	0.1481	18	0.0091	
19	-0.0199		24.9411	0.1625	19	-0.0249	
20	0.0084		25.0706	0.1987	20	0.0021	
21	0.0257		26.2890	0.1956	21	0.0251	
22	-0.0071		26.3830	0.2356	22	-0.0095	
23	0.0001		26.3830	0.2831	23	-0.0065	
24	0.0084		26.5124	0.3277	24	-0.0004	
25	-0.0058		26.5735	0.3775	25	-0.0140	

Humidity: Seasonal ARIMA(2,0,1)(0,1,1)365 without intercept

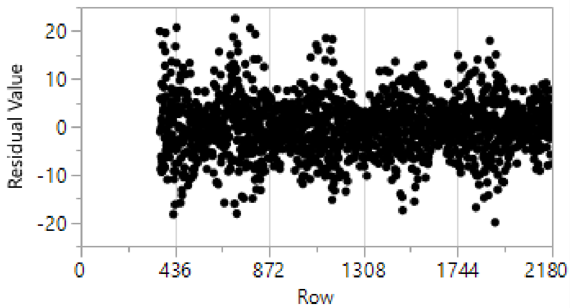
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	1.389383	0.0323433	42.96	<.0001*
AR1,2	1	2	-0.399263	0.0275574	-14.49	<.0001*
MA1,1	1	1	0.954219	0.0210019	45.43	<.0001*
MA2,365	2	365	1.000000	0.1105408	9.05	<.0001*



Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	-0.0008		0.0012	0.9721	1	-0.0008	
2	-0.0042		0.0338	0.9832	2	-0.0042	
3	-0.0059		0.0973	0.9922	3	-0.0059	
4	0.0011		0.0995	0.9988	4	0.0011	
5	0.0182		0.7015	0.9829	5	0.0181	
6	-0.0108		0.9140	0.9887	6	-0.0108	
7	-0.0098		1.0899	0.9932	7	-0.0097	
8	0.0037		1.1143	0.9974	8	0.0038	
9	-0.0196		1.8177	0.9940	9	-0.0199	
10	-0.0234		2.8193	0.9854	10	-0.0239	
11	0.0277		4.2252	0.9629	11	0.0280	
12	0.0307		5.9556	0.9183	12	0.0307	
13	0.0071		6.0476	0.9444	13	0.0068	
14	-0.0095		6.2129	0.9608	14	-0.0082	
15	0.0058		6.2747	0.9748	15	0.0067	
16	0.0190		6.9343	0.9745	16	0.0170	
17	-0.0064		7.0093	0.9834	17	-0.0073	
18	-0.0128		7.3122	0.9872	18	-0.0120	
19	0.0126		7.6048	0.9902	19	0.0128	
20	-0.0337		9.6906	0.9734	20	-0.0339	
21	-0.0154		10.1250	0.9772	21	-0.0139	
22	0.0010		10.1268	0.9851	22	0.0027	
23	-0.0119		10.3893	0.9886	23	-0.0136	
24	-0.0160		10.8618	0.9900	24	-0.0188	
25	-0.0442		14.4678	0.9530	25	-0.0422	

Pressure: Seasonal ARIMA(3,0,0)(0,1,1)365 without intercept

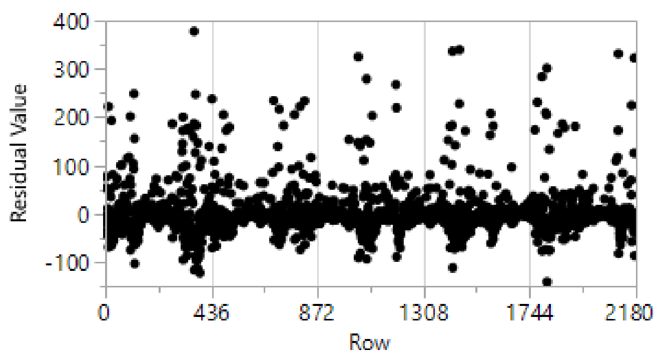
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	0.6120875	0.0235538	25.99	<.0001*
AR1,2	1	2	-0.1284732	0.0273676	-4.69	<.0001*
AR1,3	1	3	0.0527441	0.0234562	2.25	0.0247*
MA2,365	2	365	0.8606437	0.0818640	10.51	<.0001*



Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	0.0042		0.0320	0.8581	1	0.0042	
2	-0.0021		0.0397	0.9804	2	-0.0021	
3	0.0264		1.3125	0.7262	3	0.0265	
4	-0.0363		3.7073	0.4471	4	-0.0365	
5	0.0199		4.4293	0.4894	5	0.0204	
6	0.0082		4.5509	0.6026	6	0.0071	
7	0.0076		4.6568	0.7018	7	0.0096	
8	0.0659		12.5859	0.1269	8	0.0636	
9	0.0165		13.0860	0.1588	9	0.0171	
10	0.0246		14.1904	0.1645	10	0.0247	
11	0.0090		14.3381	0.2148	11	0.0059	
12	0.0113		14.5700	0.2658	12	0.0149	
13	-0.0088		14.7128	0.3256	13	-0.0116	
14	-0.0029		14.7281	0.3970	14	-0.0031	
15	0.0170		15.2563	0.4331	15	0.0146	
16	-0.0033		15.2761	0.5045	16	-0.0072	
17	0.0278		16.6982	0.4750	17	0.0245	
18	-0.0058		16.7608	0.5396	18	-0.0106	
19	0.0251		17.9145	0.5282	19	0.0253	
20	-0.0078		18.0266	0.5857	20	-0.0128	
21	0.0072		18.1218	0.6413	21	0.0105	
22	0.0175		18.6825	0.6648	22	0.0145	
23	0.0330		20.6912	0.5999	23	0.0339	
24	-0.0199		21.4224	0.6137	24	-0.0223	
25	-0.0057		21.4815	0.6655	25	-0.0090	

Cumulated wind speed: ARIMA(20,1,0) without intercept

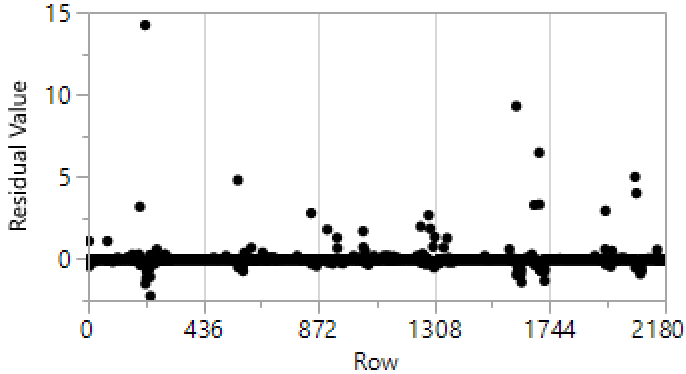
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	-0.7907894	0.0213869	-36.98	<.0001*
AR1,2	1	2	-0.7286888	0.0272446	-26.75	<.0001*
AR1,3	1	3	-0.6922830	0.0313267	-22.10	<.0001*
AR1,4	1	4	-0.6857445	0.0344951	-19.88	<.0001*
AR1,5	1	5	-0.6572678	0.0372872	-17.63	<.0001*
AR1,6	1	6	-0.5804047	0.0395651	-14.67	<.0001*
AR1,7	1	7	-0.5847754	0.0410150	-14.26	<.0001*
AR1,8	1	8	-0.5495578	0.0422550	-13.01	<.0001*
AR1,9	1	9	-0.4264688	0.0432612	-9.86	<.0001*
AR1,10	1	10	-0.4105155	0.0435083	-9.44	<.0001*
AR1,11	1	11	-0.3720290	0.0435267	-8.55	<.0001*
AR1,12	1	12	-0.3233225	0.0432952	-7.47	<.0001*
AR1,13	1	13	-0.3355708	0.0423245	-7.93	<.0001*
AR1,14	1	14	-0.2904030	0.0411231	-7.06	<.0001*
AR1,15	1	15	-0.2274383	0.0397661	-5.72	<.0001*
AR1,16	1	16	-0.1991267	0.0376367	-5.29	<.0001*
AR1,17	1	17	-0.1639447	0.0350354	-4.68	<.0001*
AR1,18	1	18	-0.1009037	0.0318811	-3.17	0.0016*
AR1,19	1	19	-0.0626110	0.0276198	-2.27	0.0235*
AR1,20	1	20	-0.0448360	0.0214961	-2.09	0.0371*



Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	-0.0007		0.0009	0.9755	1	-0.0007	
2	-0.0021		0.0105	0.9948	2	-0.0021	
3	-0.0042		0.0499	0.9971	3	-0.0042	
4	-0.0052		0.1089	0.9986	4	-0.0052	
5	-0.0066		0.2037	0.9991	5	-0.0066	
6	-0.0090		0.3825	0.9990	6	-0.0091	
7	-0.0102		0.6116	0.9989	7	-0.0103	
8	-0.0122		0.9353	0.9986	8	-0.0123	
9	-0.0144		1.3881	0.9979	9	-0.0146	
10	-0.0157		1.9260	0.9969	10	-0.0160	
11	-0.0132		2.3093	0.9971	11	-0.0137	
12	-0.0170		2.9443	0.9959	12	-0.0176	
13	-0.0216		3.9723	0.9915	13	-0.0225	
14	-0.0215		4.9847	0.9860	14	-0.0225	
15	-0.0248		6.3374	0.9735	15	-0.0262	
16	-0.0236		7.5606	0.9609	16	-0.0254	
17	-0.0187		8.3261	0.9591	17	-0.0208	
18	-0.0250		9.6998	0.9413	18	-0.0276	
19	-0.0312		11.8366	0.8925	19	-0.0344	
20	-0.0285		13.6244	0.8490	20	-0.0324	
21	-0.0301		15.6141	0.7909	21	-0.0348	
22	-0.0308		17.7058	0.7233	22	-0.0365	
23	-0.0075		17.8313	0.7666	23	-0.0140	
24	0.0316		20.0385	0.6946	24	0.0250	
25	0.0031		20.0590	0.7437	25	-0.0037	

Precipitation: ARIMA(19,1,0) without intercept:

Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	-0.8981154	0.0211674	-42.43	<.0001*
AR1,2	1	2	-0.8694089	0.0282086	-30.82	<.0001*
AR1,3	1	3	-0.8425130	0.0334101	-25.22	<.0001*
AR1,4	1	4	-0.8072236	0.0372660	-21.66	<.0001*
AR1,5	1	5	-0.7290874	0.0403028	-18.09	<.0001*
AR1,6	1	6	-0.6784610	0.0423589	-16.02	<.0001*
AR1,7	1	7	-0.6443339	0.0438106	-14.71	<.0001*
AR1,8	1	8	-0.6098031	0.0448076	-13.61	<.0001*
AR1,9	1	9	-0.5750765	0.0454170	-12.66	<.0001*
AR1,10	1	10	-0.5384703	0.0456010	-11.81	<.0001*
AR1,11	1	11	-0.4995184	0.0453696	-11.01	<.0001*
AR1,12	1	12	-0.4701866	0.0447261	-10.51	<.0001*
AR1,13	1	13	-0.4315853	0.0437070	-9.87	<.0001*
AR1,14	1	14	-0.4001418	0.0422366	-9.47	<.0001*
AR1,15	1	15	-0.3709589	0.0401660	-9.24	<.0001*
AR1,16	1	16	-0.3397096	0.0371210	-9.15	<.0001*
AR1,17	1	17	-0.2409943	0.0332673	-7.24	<.0001*
AR1,18	1	18	-0.2191059	0.0280824	-7.80	<.0001*
AR1,19	1	19	-0.1497067	0.0210758	-7.10	<.0001*



Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
1	-0.0026		0.0150	0.9025	1	-0.0026	
2	-0.0056		0.0843	0.9587	2	-0.0056	
3	-0.0091		0.2662	0.9663	3	-0.0092	
4	-0.0108		0.5229	0.9712	4	-0.0109	
5	-0.0110		0.7897	0.9777	5	-0.0112	
6	-0.0145		1.2513	0.9743	6	-0.0148	
7	-0.0150		1.7464	0.9725	7	-0.0155	
8	-0.0162		2.3214	0.9695	8	-0.0168	
9	-0.0181		3.0364	0.9628	9	-0.0189	
10	-0.0195		3.8665	0.9532	10	-0.0206	
11	-0.0201		4.7565	0.9424	11	-0.0216	
12	-0.0210		5.7260	0.9293	12	-0.0228	
13	-0.0224		6.8273	0.9108	13	-0.0246	
14	-0.0235		8.0427	0.8871	14	-0.0263	
15	-0.0252		9.4375	0.8536	15	-0.0285	
16	-0.0270		11.0345	0.8074	16	-0.0309	
17	-0.0289		12.8757	0.7445	17	-0.0337	
18	-0.0310		14.9963	0.6622	18	-0.0367	
19	-0.0299		16.9601	0.5926	19	-0.0366	
20	-0.0304		19.0019	0.5217	20	-0.0384	
21	-0.0213		19.9973	0.5214	21	-0.0304	
22	-0.0077		20.1271	0.5751	22	-0.0178	
23	0.0074		20.2490	0.6268	23	-0.0033	
24	-0.0022		20.2594	0.6819	24	-0.0132	
25	-0.0119		20.5728	0.7162	25	-0.0234	

Based on these ARIMA models, both the related out-of-sample predictors and the out-of-sample PM 2.5 could be forecasted. I then compared these predictions with the real out-of-sample observations to derive the MSE:

ARMA(1,1) prediction	ARMA(1,1) MSE		ARMA(2,2) prediction	ARMA(2,2) MSE
107.103	40972.7		109.254	41048.9
57.799			56.171	
69.142			70.978	
57.412			56.168	
76.451			78.107	
71.835			70.477	
85.743			87.253	
65.312			63.956	
53.444			55.084	
50.588			49.583	

ARMA (1,1) has a smaller MSE than ARMA (2, 2), and ARMA (1,1) is less complicated than ARMA (2,2). Thus, ARIMA (1,1,1) was picked as the best model for the autocorrelated regression noise.

The explicit model equation is given below:

$$PM_t' = 4.9951DEWP_t' + 0.9930HUMI_t' - 1.2625PRES_t' - 0.0789lws_t' - 8.8041precip_t' + N_t'$$

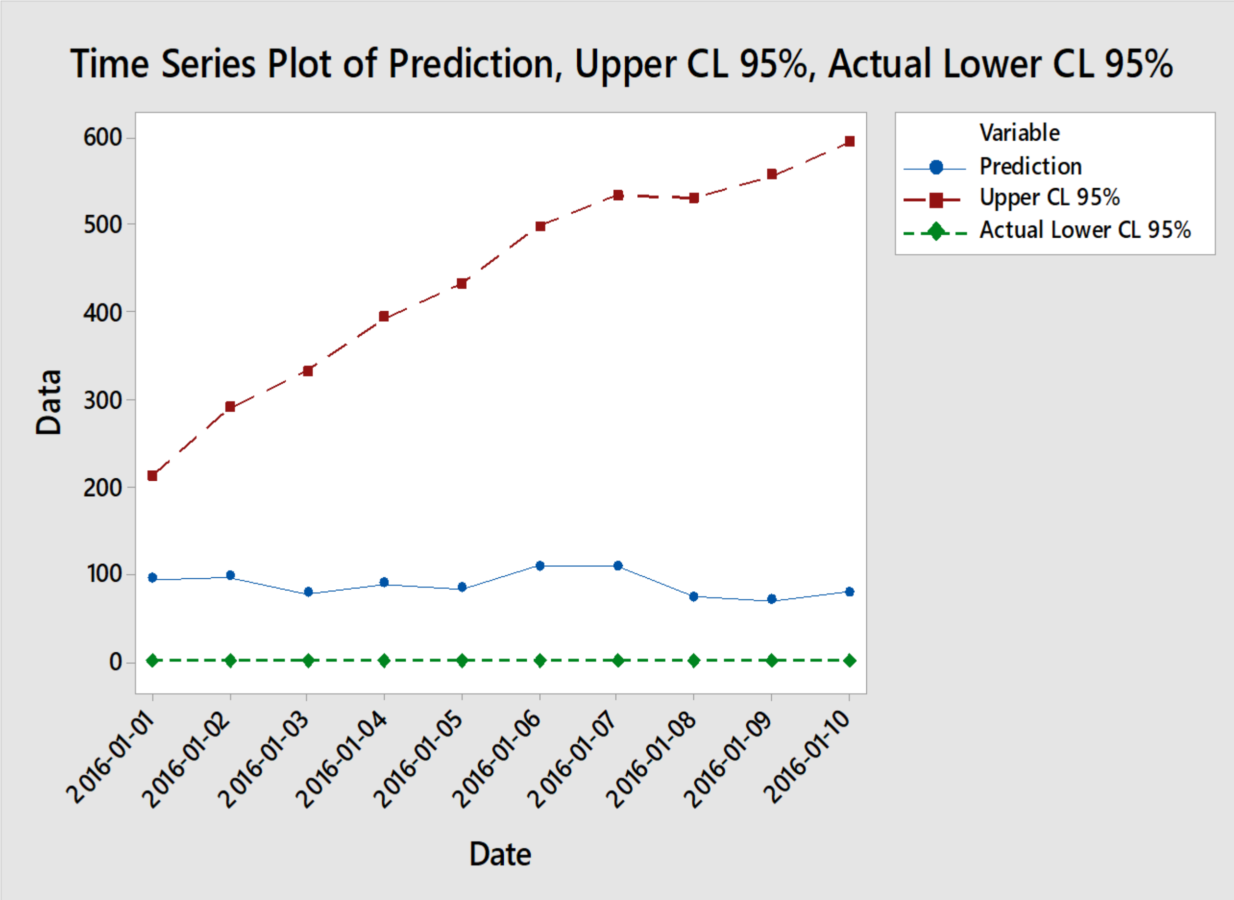
$$(1-B)(1-0.3170B)N_t = (1-0.8893B)e_t$$

$$N_t = 1.3170N_{t-1} - 0.3170N_{t-2} + e_t - 0.8893e_{t-1}$$

After validating these models, I forecasted the predictors of the next ten days. Then, I predicted the future ten days' PM2.5 values based on these assumptive predictors.

Date	dewp10	humi10	pres10	lws10	preci10	Prediction	Upper CL 95%	Lower CL 95%
2016-01-01	-11.7451	50.3034	1030.63	23.2697	0	94.501	212.210	-23.208
2016-01-02	-14.7791	53.3397	1031.04	32.1560	0	96.505	290.109	-97.099
2016-01-03	-17.8006	43.7258	1029.70	42.9152	0	77.181	331.377	-177.015
2016-01-04	-17.3853	47.5155	1027.52	33.9663	0	88.464	393.209	-216.281
2016-01-05	-16.8872	43.1998	1029.84	28.3457	0	83.661	432.188	-264.866
2016-01-06	-14.9548	60.1765	1030.80	33.9203	0	109.575	497.104	-277.954
2016-01-07	-14.6486	53.9715	1028.30	17.9857	0	109.172	532.160	-313.817
2016-01-08	-18.4927	43.2624	1031.95	16.9652	0	74.304	530.012	-381.404
2016-01-09	-19.1638	40.2200	1030.71	15.7897	0	69.697	555.931	-416.537
2016-01-10	-17.9988	41.8460	1029.14	15.7224	0	79.193	594.147	-435.762

Since PM 2.5 cannot be negative, the lower 95% confidence bound was set as 0. The time series plot for the PM 2.5 predictions in the next ten days（2016-01-01 to 2016-01-10） is shown below:


Time Series Plot of Prediction, Upper CL 95%, Actual Lower CL 95%
Data
600
500
400
300
200
100
0
2016-01-01
2016-01-02
2016-01-03
2016-01-04
2016-01-05
2016-01-06
2016-01-07
2016-01-08
2016-01-09
2016-01-10
Date
Variable
Prediction
Upper CL 95%
Actual Lower CL 95%

Conclusion:

Considering even the best model for daily forecast has an MSE larger than 40,000, it is challenging to predict the daily PM 2.5 in the future accurately. The later the predicted date is, the further the upper 95% confidence limit expands. Therefore, it is impossible to forecast Beijing's long-term daily air pollution level, as the 95% confidence interval will only become larger and larger. However, in the short term, the daily predictions are still worthwhile.

On the other hand, because the monthly forecast model focuses on the average PM 2.5 of each month themselves and has excluded all independent variables, it has a comparatively smaller MSE (1704.62). Therefore, the predictions of the monthly average PM 2.5 are much more precise than the daily PM 2.5 forecasts, in both the short and long term.

According to the daily forecasts, Beijing's PM 2.5 would most likely be around 100 in the first ten days of 2016. People shall stay indoors if possible and wear masks while working outdoors.

However, as the monthly forecasts show, Beijing's air quality will likely improve after February 2016 as March's average PM 2.5 will drop about twenty percent compared to February's average PM 2.5. Thus, I suggest Beijing residents practice outdoor activities in Spring 2016.

Based on my daily forecast model, precipitation has the most negative impact on PM 2.5 compared to the other predictors. This relationship is reasonable because rain can lower PM 2.5 by taking away the particulates in the air. Therefore, I recommend the Beijing government to create artificial rainfall when Beijing's air pollution reaches a hazardous level. However, because artificial rain may have a potential negative impact on the macro environment of China, the most critical objective of Beijing is to set an environmentally friendly development scheme instead of the extensive growth it adopts now. It will surely bring temporary pain, but in the long term, it is the only way to fix Beijing's horrible air pollution completely.

Reference:

S. X. Chen (2016). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities]. Beijing, China: Peking University, Guanghua School of Management, Center for Statistical Science.