Time Series Analysis of PM2.5 In Beijing

DNSC 6219: Time Series Forecasting

Team Members:

Pei-Hsuan Hsia

Jingbo Zhang

Agnes Jiang

Jiwei Zeng

May 15, 2018

1. Introduction and Overview

Our dataset is found at https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data.

The daily dataset (a subset from the original one) we use contains the PM2.5 data of US Embassy in Beijing at 10 a.m. from Jan 1st, 2013 to Dec 31st, 2014. For the total of 730 observations, 7 of the PM2.5 data points are missing. We used the closest available data after that point to fill the missing. Meanwhile, meteorological data from Beijing Capital International Airport are also included. The independent variables are shown in table 1.1.

Variable	Description
DEWP	Dew Point (℃)
TEMP	Temperature (℃)
PRES	Pressure (hPa)
cbwd	Combined wind direction
lws	Cumulated wind speed (m/s)
ls	Cumulated hours of snow
lr	Cumulated hours of rain

Table 1.1 independent variable list

In our analysis, we will treat PM2.5 as our dependent variable (Y) and select some of the meteorological data as independent variables (X).

From the series plot (figure 1.1), we don't observe obvious trends over time, but we can see higher values and higher variance in the beginning months of each year. Since there is high variance in the original series, we think a square root transformation is necessary to stabilize the variance across time. The transformed series (figure 1.2) has more constant variability across time, therefore we will use the square root transformation in all the models.

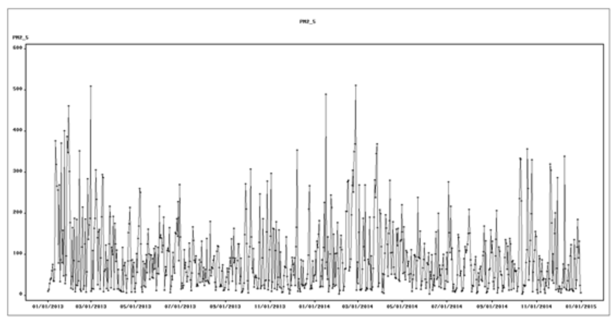

Figure 1.1 Daily PM2.5 at US Embassy in Beijing (01/01/2013-12/31/2014)

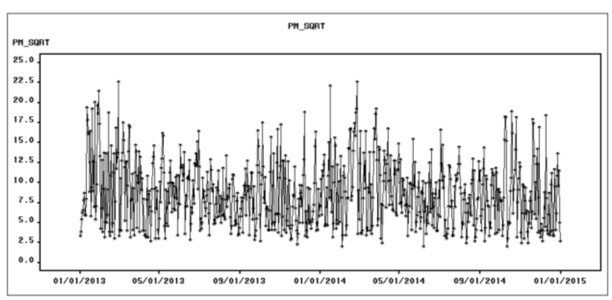

Figure 1.2 Square Root Transformed PM2.5

From the monthly box plots (figure 1.3), we see different patterns in different months. In January, February and March, we have much higher average PM2.5 values. In August we have lowest average PM2.5. Therefore, we think there is apparent seasonal behavior in different months.

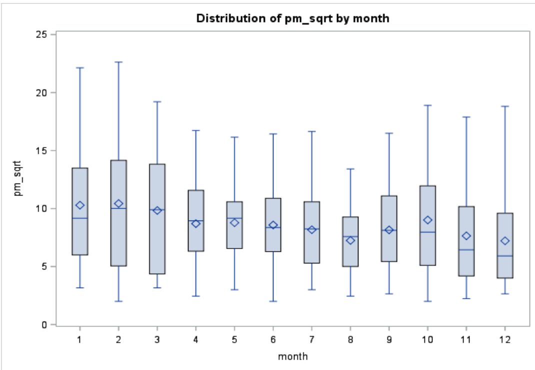

Figure 1.3 Monthly box plots

From the box plot of weekdays (figure 1.4), we can see that the means of different week days remain the same value. Therefore, we will not use week day dummy variables.

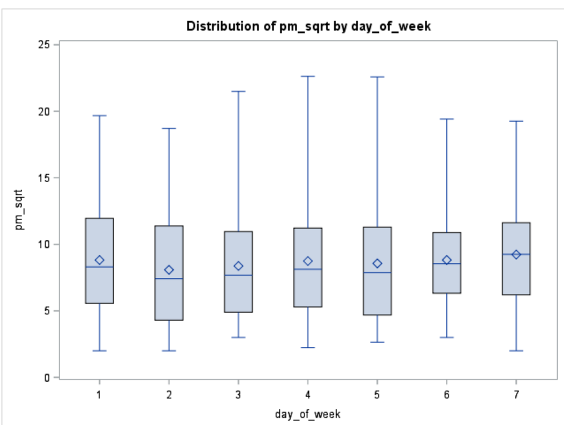

Figure 1.4 Day of week box plots

From the sample autocorrelation plot (figure 1.5), we can find that the both ACF and PACF decay quickly, which implies that the series may be described by an ARMA model.

For the following analysis, we will use 61 hold-out sample, which is the last two months in the series.

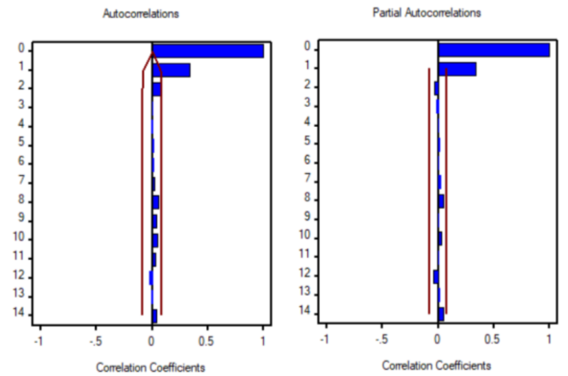

Figure 1.5 Sample autocorrelation

2. Univariate Time-series models

2.1 Deterministic Models (Seasonal Dummies) and Error model

We first fit a monthly dummies and linear trend model. But as figure 2.1 shows, the p-value associating to linear trend is 0.798 which is not significant. We remove linear trend and fit a new model.

Parameter Estimates

PM2_5

Sqrt JAN + FEB + MAR + APR + MAY + JUN + JUL + AUG + SEP + OCT + NOV + Linear Trend

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	7.57857	0.7925	9.5624	<.0001
JAN	2.66380	0.9013	2.9557	0.0048
FEB	2.80331	0.9133	3.0695	0.0035
MAR	2.19487	0.8948	2.4530	0.0179
APR	1.05456	0.8976	1.1749	0.2458
MAY	1.12364	0.8914	1.2606	0.2135
JUN	0.92259	0.8958	1.0299	0.3082
JUL	0.50610	0.8913	0.5678	0.5728
AUG	-0.42582	0.8926	-0.4771	0.6355
SEP	0.48010	0.8996	0.5337	0.5960
OCT	1.32351	0.8976	1.4745	0.1469
NOV	-0.31684	1.0376	-0.3054	0.7614
Linear Trend	0.0002317	0.000899	0.2578	0.7977
Model Variance (sigma squared)	16.40334	.	.	.

Figure 2.1 Parameter Estimation for Seasonal Dummies and Trend Model

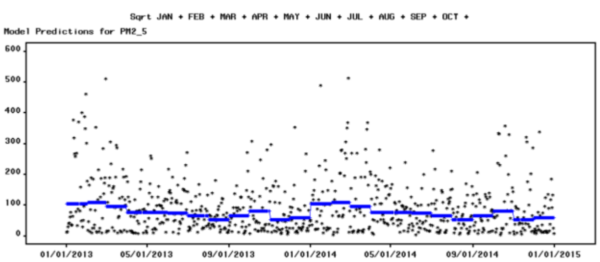

Figure 2.2 Actual vs. fitted values of Seasonal Dummies Model

For the seasonal dummies model, as we can see from the series plot with fitted values (figure 2.2), it does not have a good fit. From the ACF and PACF results of residuals (figure 2.3), we would suggest to have an error model on AR(1). The result of Seasonal Dummies Model with AR(1) error model is shown in figure 2.4.

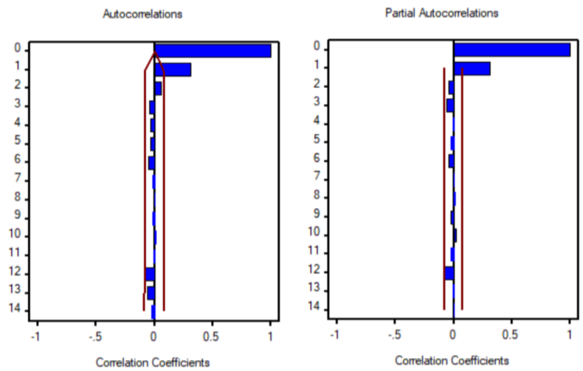

Figure 2.3 ACF and PACF of the residuals for Seasonal Dummies Model

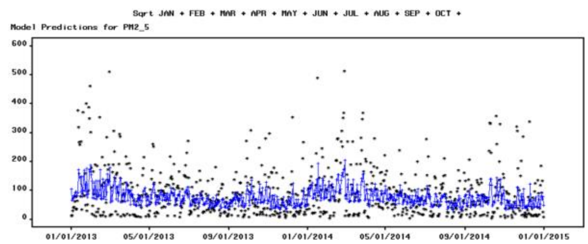

Figure 2.4 Actual vs. fitted values of Seasonal Dummies Model with AR(1) error model

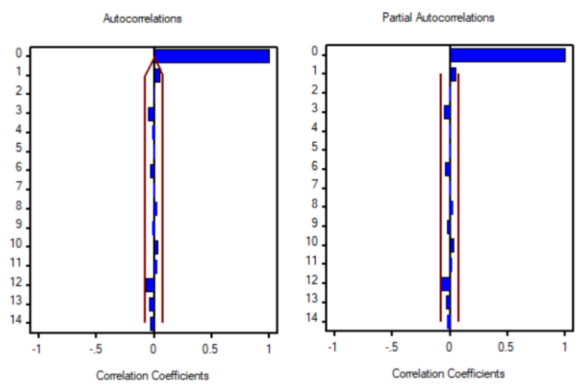

Figure 2.5 ACF and PACF of the residuals for Seasonal Dummies Model with AR(1) error model

The plot of ACF and PACF of residuals (figure 2.5) shows that we cannot reject the hypothesis that residual ACFs are white noise. According to the parameter estimates (figure 2.6), only January and February are significantly different than December, our reference month.

Sqrt JAN + FEB + MAR + APR + MAY + JUN + JUL + AUG + SEP + OCT + NOV + AR(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	7.79054	0.9772	7.9723	<.0001
Autoregressive, Lag 1	0.30448	0.0372	8.1778	<.0001
JAN	2.49188	1.1895	2.0949	0.0415
FEB	2.80772	1.2169	2.3072	0.0254
MAR	1.91173	1.1969	1.5972	0.1168
APR	0.82755	1.2034	0.6877	0.4950
MAY	1.06065	1.1968	0.8862	0.3799
JUN	0.72534	1.2032	0.6028	0.5495
JUL	0.36442	1.1969	0.3045	0.7621
AUG	-0.54771	1.1969	-0.4576	0.6493
SEP	0.40097	1.2032	0.3333	0.7404
OCT	1.18817	1.1972	0.9925	0.3260
NOV	-0.47667	1.3792	-0.3456	0.7311
Model Variance (sigma squared)	14.89365	.	.	.

Figure 2.6 Parameter Estimation for Seasonal Dummies and Trend with AR(1) error model

2.2 Simple Exponential Smoothing Model

As we can see from the series plot (figure 2.7), Simple Exponential Smoothing Model does not have a good fit.

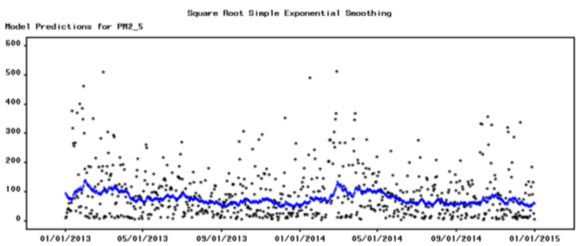

Figure 2.7 Actual versus fitted values of Simple Exponential Smoothing Model

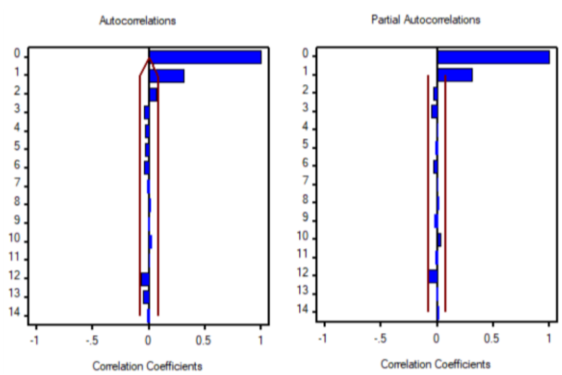

Figure 2.8 ACF and PACF of the residuals for Simple Exponential Smoothing Model

From the plot of ACF and PACF of residuals (figure 2.8), it clearly indicates to reasonably reject the hypothesis that residual ACFs are white noise, showing a not good model as well, which doesn't show a good fit neither.

From figure 2.9, we can see the level trend is significant, and residual variance is 16.767.

We also tried linear ES model, but the error is larger than simple ES and trend is not significant. Hence, we only include simple ES model in this part.

Parameter Estimates
PM2_5
Square Root Simple Exponential Smoothing

Model Parameter	Estimate	Std. Error	T	Prob>|T|
LEVEL Smoothing Weight	0.04371	0.0081	5.3844	<.0001
Residual Variance (sigma squared)	16.76660	.	.	.
Smoothed Level	9.56012	.	.	.

Figure 2.9 Parameter Estimation for the Linear Exponential Smoothing Model

2.3 ARIMA models

Since the ACF of original series decays exponentially and the PACF is chopped off after lag1, we can use AR(1) to model the series. Following are the outputs.

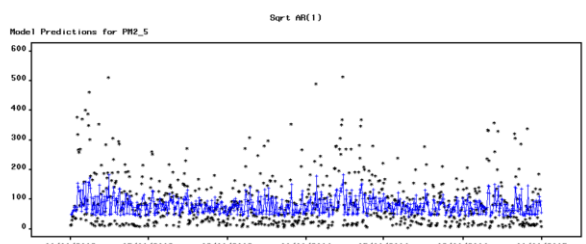

Figure 2.10 Actual versus fitted values for AR(1) model

The series plot with fitted values (figure 2.10) show that AR(1) Model cannot fit values with high variance.

From figures 2.11, we can find that the residual of AR(1)is stationary and looks like white noise. But the AR(1) model cannot capture the high variance of original series either.

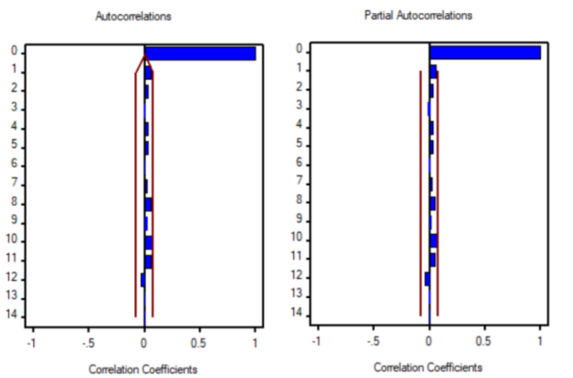

Figure 2.11 ACF and PACF of the residuals for AR(1) Model

Figure 2.12 shows the estimate parameters in AR(1). We will try to bring in more dummies and other regressors in next steps of the model fitting.

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	8.77979	0.2284	38.4364	<.0001
Autoregressive, Lag 1	0.34243	0.0364	9.4106	<.0001
Model Variance (sigma squared)	15.11990	.	.	.

Figure 2.12 Parameter Estimation for AR(1) Model

2.4 Comparison of models

For all the above models we've fitted, we have the same hold-out samples of 61 (2 months - November and December). Based on the root mean square error comparison in table 2.1, AR(1) model performs better in period of fit, while seasonal dummies with error model performs better in hold-out sample.

Model	Hold-out (RMSE)	Period of Fit (RMSE)
Seasonal dummies	81.911	4.047
Seasonal dummies with error model	79.352	3.859
Simple Exponential Smoothing	82.239	4.095
AR(1)	78.423	3.888

Table 2.1 Comparison of 4 Univariate Time-series models in RMSE

3. Multivariate Time Series Models

3.1 First Independent Variable: PRES

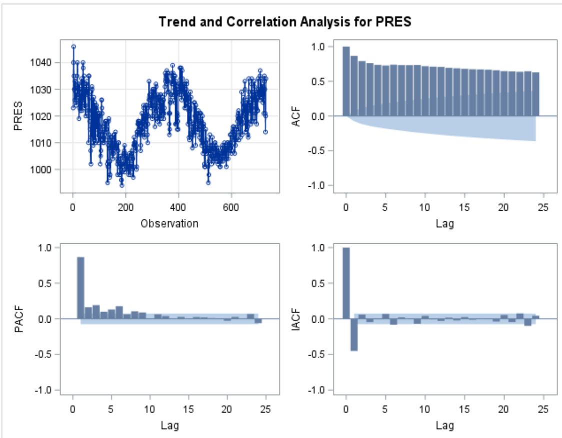

Figure 3.1 Correlation Analysis for PRES

As shown in figure 3.1, the series of PRES is not stationary, therefore we need to take the difference first.

From figure 3.2, we find that PRES(1) is not white noise and it looks like MA(2). After fitting it with MA(2) model, it doesn't pass the white noise test.

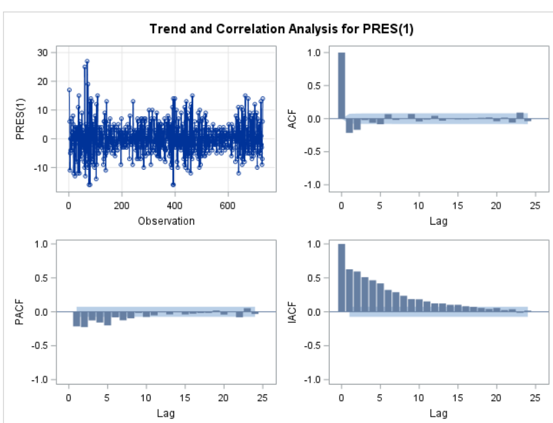

Figure 3.2 Correlation analysis for PRES(1)

We tried another model ARMA(2, 2) and it passes white noise test. As in figure 3.3, after fitting it with ARIMA(2, 1, 2) error model, the ACF and PACF of residual is not significant than 0, and the residual passes the white noise test, which indicates the residual is white noise and the PRES series has been pre-whitened.

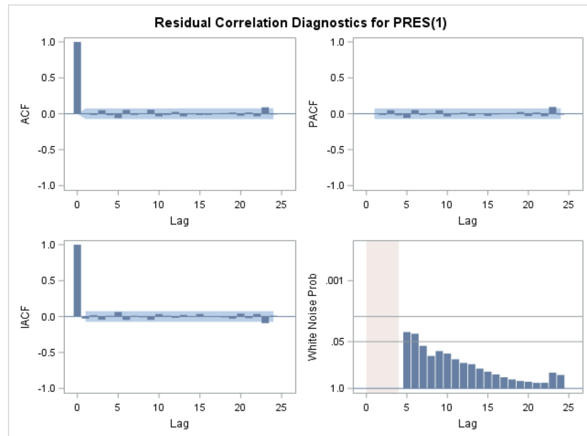

Figure 3.3 Residual Correlation Diagnostics for PRES(1)

Therefore, we used the cross correlation result associating with ARIMA(2, 1, 2) to identify the transfer function model.

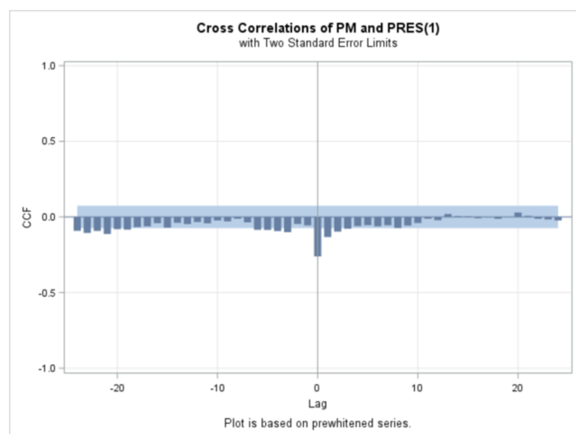

Figure 3.4 Cross Correlation of sqrt(PM2.5) and Differenced PRES

From the cross correlation between sqrt(PM2.5) and first difference of PRES (figure 3.4), there is no response until lag 0 and the cross-correlation decays exponentially starting at lag 0, which indicates that we should set b=0, s=0, r=1 for the first difference of PRES.

After fitting the TF model, the residual correlation is shown in figure 3.5. Obviously, the residual for TF model is not white noise. From the ACF and PACF, we would suggest to have an error model on AR (1).

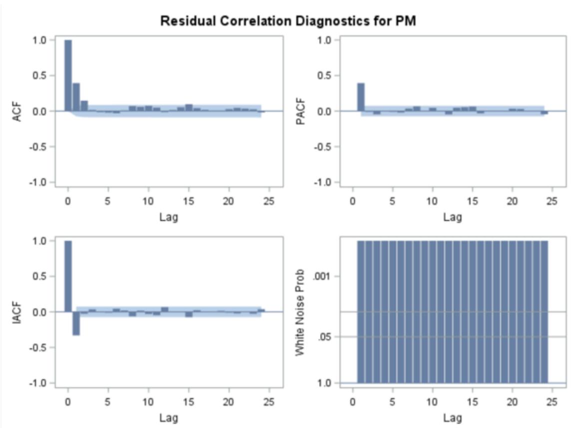

Figure 3.5 Residual correlation for TF model on PRES

The autocorrelation check of residual after fitting an error model is shown in figure 3.6. We can find that the residual is white noise now since the p value is high.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.50	5	0.7760	0.005	0.010	-0.041	-0.017	-0.009	-0.035
12	10.06	11	0.5250	-0.002	0.062	0.014	0.048	0.038	-0.048
18	16.10	17	0.5170	0.002	0.022	0.087	0.004	0.001	-0.001
24	17.98	23	0.7584	-0.006	0.013	0.033	0.016	0.028	-0.014
30	39.07	29	0.1003	-0.025	-0.073	-0.067	0.100	0.067	0.054
36	42.10	35	0.1905	0.030	0.036	0.023	0.008	-0.031	0.014
42	48.30	41	0.2016	0.013	0.011	0.054	0.006	0.058	-0.038
48	53.57	47	0.2368	0.052	-0.003	-0.005	0.059	-0.006	0.020

Figure 3.6 Autocorrelation check of residuals

The cross-correlation check between residuals and PRES is shown in figure 3.7. We can conclude that the model is appropriate because there's no cross correlation between residual and PRES.

Crosscorrelation Check of Residuals with Input PRES									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	1.72	5	0.8858	-0.001	0.009	-0.037	0.022	0.007	-0.020
11	6.95	11	0.8035	0.016	-0.002	-0.072	0.009	-0.027	0.030
17	12.97	17	0.7384	-0.030	0.075	-0.021	0.004	-0.036	-0.006
23	19.64	23	0.6636	-0.026	0.015	0.069	-0.015	-0.051	0.024
29	28.87	29	0.4721	-0.062	0.011	-0.011	0.085	-0.035	0.013
35	34.60	35	0.4872	-0.028	-0.048	-0.012	0.006	0.055	-0.040
41	35.63	41	0.7078	0.012	0.010	-0.018	0.008	-0.011	-0.026
47	39.51	47	0.7729	0.026	-0.005	0.057	-0.004	-0.026	0.026

Figure 3.7 Cross correlation check of residuals with PRES

The estimated parameters are shown in figure 3.8. We can find that all the parameters are significant.

PM2_5
Sqrt PRES[Dif(1) / D(1)] + AR(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	8.67621	0.2155	40.2632	<.0001
Autoregressive, Lag 1	0.39439	0.0342	11.5469	<.0001
PRES[Dif(1) / D(1)]	-0.32341	0.0255	-12.6596	<.0001
PRES[Dif(1) / D(1)] Den1	0.38820	0.0799	4.8571	<.0001
Model Variance (sigma squared)	12.42093	.	.	.

Figure 3.8 Estimated parameters for TF mode on PRES

3.2 Independent Model: Iws

From figure 3.9, we can see that Iws is stationary. Therefore, we don't need to take difference. The Iws series looks like an AR(1) model. To pre-whiten the series of Iws, we use AR(1) model to fit it.

After fitting it with AR(1) model, the ACF and PACF of residual is not significant than 0 (figure 3.10), and the residual passes the white noise test, which indicates the residual is white noise and the model performs well.

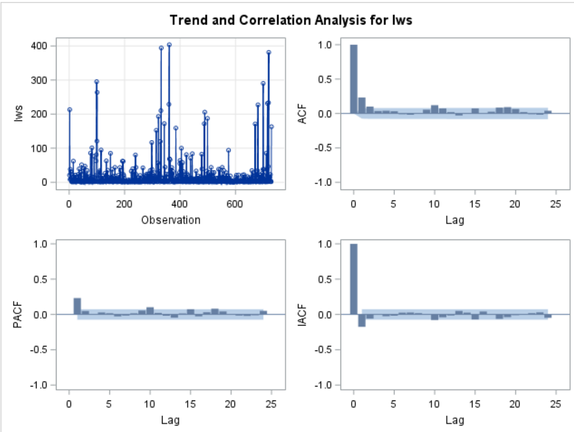

Figure 3.9 Correlation Analysis for Iws

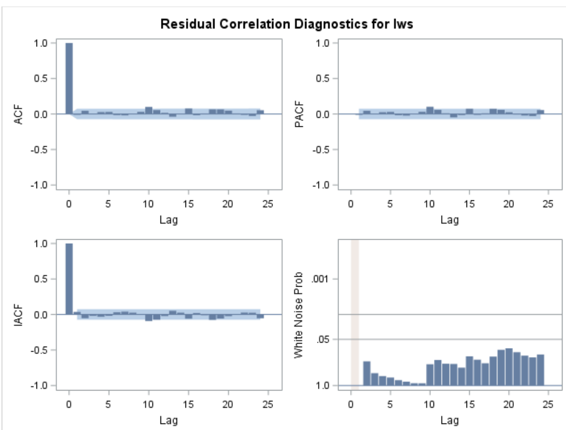

Figure 3.10 Residual Correlation Diagnostics for Iws

From the cross-correlation plot between PM2.5 and Iws (figure 3.11), there is no response until lag 0 and the cross-correlation exponential decays starting at lag 0, which indicates that we should set b = 0, s =0, r = 1.

After fitting the TF model, the residual correlation is shown in figure 3.12. Obviously, the residual for TF model is not white noise. From the ACF and PACF, we would suggest to have an error model on AR (1).

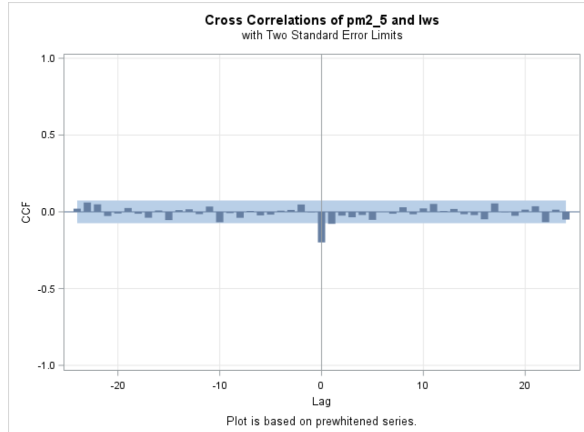

Figure 3.11 Cross Correlation of PM2.5 and Iws

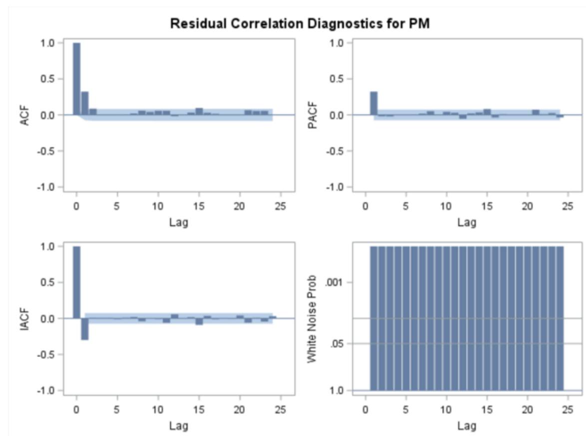

Figure 3.12 Residual correlation for TF model on Iws

The autocorrelation check of residual after error model is shown in figure 3.13. We can find that the residual is white noise now since the p value is high. The cross-correlation check between residuals and Iws is shown in figure 3.14. We can conclude that the model is appropriate because there's no cross correlation between residual and Iws.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.63	5	0.9868	0.006	-0.010	-0.026	-0.006	-0.001	0.000
12	7.02	11	0.7976	0.002	0.050	0.009	0.034	0.056	-0.042
18	14.03	17	0.6647	-0.006	0.006	0.096	-0.004	0.008	-0.001
24	19.78	23	0.6552	0.002	-0.028	0.062	0.024	0.048	-0.012
30	36.34	29	0.1638	-0.016	-0.036	-0.015	0.124	0.044	0.052
36	40.44	35	0.2424	0.041	0.047	-0.010	0.011	-0.028	0.023
42	46.45	41	0.2578	0.054	0.040	0.022	0.019	0.040	-0.029
48	54.43	47	0.2127	0.052	-0.031	-0.002	0.074	-0.003	0.033

Figure 3.13 Autocorrelation check of residuals for TF model on Iws after error model

Crosscorrelation Check of Residuals with Input Iws									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	1.35	5	0.9296	-0.000	-0.001	0.011	-0.029	-0.011	-0.028
11	7.87	11	0.7251	0.012	-0.009	0.026	0.008	0.055	0.070
17	12.04	17	0.7979	0.002	0.006	-0.040	-0.001	-0.033	0.054
23	16.78	23	0.8201	0.015	-0.006	0.008	0.019	-0.076	0.007
29	34.95	29	0.2062	-0.036	0.026	0.073	0.126	0.035	-0.020
35	40.30	35	0.2473	-0.053	-0.044	-0.001	-0.028	-0.027	0.033
41	41.89	41	0.4321	0.019	0.029	0.004	-0.011	0.027	0.010
47	42.83	47	0.6457	0.008	-0.018	-0.011	0.002	0.007	0.027

Figure 3.14 Cross correlation check of residuals with Iws

The estimated parameters are shown in figure 3.15. We can find that all of the parameters are significant.

PM2_5
Sqrt Iws[/ D(1)] + AR(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	9.26352	0.2386	38.8292	<.0001
Autoregressive, Lag 1	0.32280	0.0352	9.1821	<.0001
IWS[/ D(1)]	-0.02405	0.0033	-7.2097	<.0001
IWS[/ D(1)] Den1	0.29308	0.1252	2.3416	0.0195
Model Variance (sigma squared)	14.26984	.	.	.

Figure 3.15 Estimated parameters for TF mode on PRES

3.3 Two independent variables model

Next, we have both PRES and Iws in the TF model, and the residual correlation is shown in figure 3.16. We can find that the residual is not white noise. From the ACF and PACF, we would suggest to have an error model on AR (1).

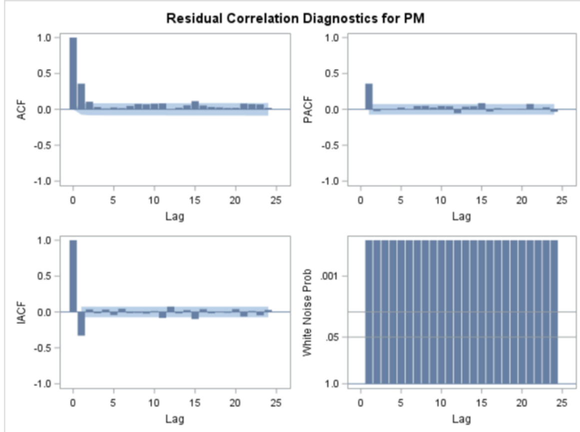

Figure 3.16 Residual correlation for TF model with two independent variables

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.33	5	0.8012	0.003	0.009	-0.041	-0.019	-0.011	-0.030
12	11.50	11	0.4026	0.009	0.064	0.011	0.052	0.048	-0.055
18	17.93	17	0.3934	0.006	0.018	0.090	0.001	0.004	-0.010
24	20.13	23	0.6342	-0.007	0.014	0.027	0.013	0.041	-0.011
30	34.88	29	0.2085	-0.019	-0.035	-0.037	0.107	0.054	0.045
36	37.88	35	0.3395	0.032	0.043	0.015	0.012	-0.020	0.017
42	42.98	41	0.3865	0.024	0.019	0.045	0.002	0.057	-0.022
48	48.87	47	0.3978	0.059	0.006	-0.005	0.058	-0.006	0.024

Figure 3.17 Autocorrelation check of residuals for TF model with two independent variables

The autocorrelation check for residuals after fitting an error model is shown in figure 3.17. We can find that the residual is white noise now.

3.4 Model comparison

From figure 3.18, we can find that model with two independent variables and AR(1) error model preforms the best among all TF models since it has the lowest root mean square error.

Forecast Model	Model Title	Root Mean Square Error
☑	Sqrt PRES[Dif(1) / D(1)] + AR(1)	74.46216
☐	Sqrt Iws[/ D(1)] + AR(1)	78.92843
☐	Sqrt PRES[Dif(1) / D(1)] + Iws[/ D(1)] + AR(1)	73.04541

Figure 3.18 Model comparison for root mean square error

4. Periodogram analysis

In this part, we will use periodogram to fit our series because there is obvious periodicity in our data.

4.1 Periodicity detection

```
data new2;
set work.new;
pm=sqrt(pm2_5);
if time<=365 then output new2;
run;

proc reg;
model pm=;
output out=new2 r=dpm;

proc spectra data=new2 p;
var dpm;

proc print;
run;
```

Figure 4.1 Code for detection

Obs	FREQ	PERIOD	P_01
1	0.00000	.	0.000
2	0.01721	365.000	174.384
3	0.03443	182.500	49.202
4	0.05164	121.667	243.212
5	0.06886	91.250	6.845
6	0.08607	73.000	23.767
7	0.10329	60.833	55.734
8	0.12050	52.143	83.424
9	0.13771	45.625	67.679
10	0.15493	40.556	88.721
11	0.17214	36.500	35.405
12	0.18936	33.182	20.414
13	0.20657	30.417	25.008
14	0.22378	28.077	23.537
15	0.24100	26.071	43.294
16	0.25821	24.333	49.275
17	0.27543	22.813	35.827
18	0.29264	21.471	1.472
19	0.30986	20.278	125.801
20	0.32707	19.211	35.299

Figure 4.2 Results (Partial)

Using proc spectra, we can detect the appropriate number of periodicity if their P_01 are larger than 100. We find out that the appropriate number includes 1, 3, 18, 20, 22, 39, 41, 48 and 71.

After detecting the number, we copy a new dataset with columns of periodicity.

```
data new3;
set sasuser.pm25;
pm=sqrt(pm2_5);
time=_N_;

IF time>365 THEN pm=.;
COS1=COS(2*3.14159*TIME*1/365);
sin1=sin(2*3.14159*TIME*1/365);

COS3=COS(2*3.14159*TIME*3/365);
sin3=sin(2*3.14159*TIME*3/365);

COS18=COS(2*3.14159*TIME*18/365);
sin18=sin(2*3.14159*TIME*18/365);

COS20=COS(2*3.14159*TIME*20/365);
sin20=sin(2*3.14159*TIME*20/365);

COS22=COS(2*3.14159*TIME*22/365);
sin22=sin(2*3.14159*TIME*22/365);

COS39=COS(2*3.14159*TIME*39/365);
sin39=sin(2*3.14159*TIME*39/365);

COS41=COS(2*3.14159*TIME*41/365);
sin41=sin(2*3.14159*TIME*41/365);

COS48=COS(2*3.14159*TIME*48/365);
sin48=sin(2*3.14159*TIME*48/365);

COS71=COS(2*3.14159*TIME*71/365);
sin71=sin(2*3.14159*TIME*71/365);
```

Figure 4.3 Creating columns of periodicity

4.2 Building models for Periodogram

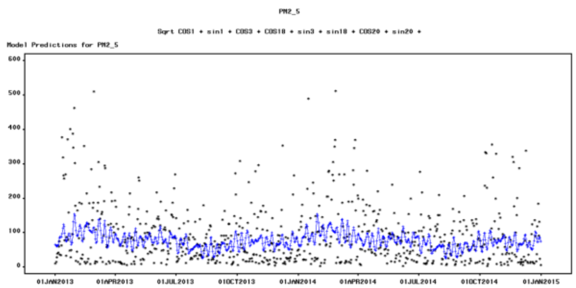

Figure 4.4 Actual versus fitted values of Periodogram Model

As we can see from the series plot (figure 4.4), periodogram model does not have a good fit.

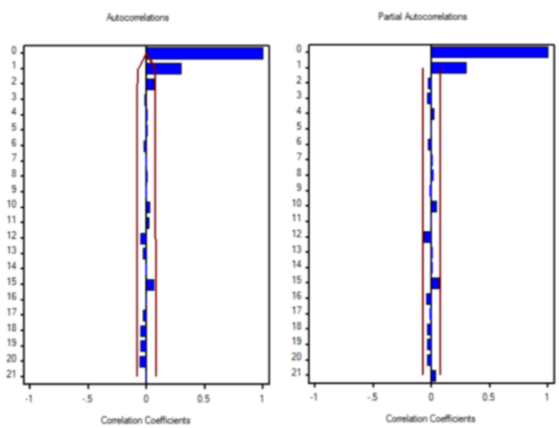

Figure 4.5 ACF and PACF of the residuals for periodogram model

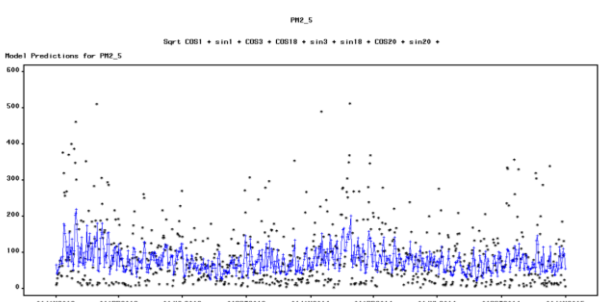

Figure 4.6 Actual versus fitted values of Periodogram Model with AR(1) error model

Obviously, the residual for Periodogram Model is not white noise (figure 4.5). From the ACF and PACF results, we would suggest to have an error model on AR(1). The result of Periodogram Model with AR(1) error model is shown in figure 4.6. The series plot doesn't show a good fit either, but it's a little better than the one without error model.

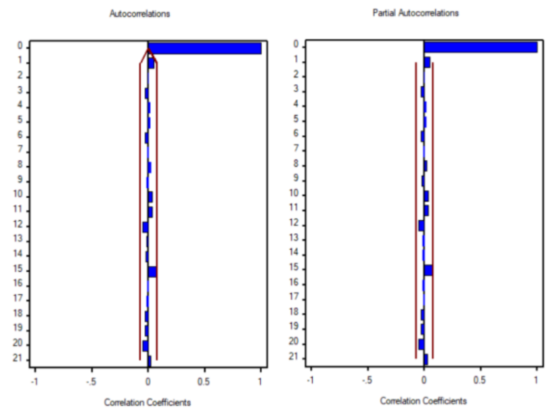

Figure 4.7 ACF and PACF of the residuals for Periodogram Model with AR(1) error model

The result of ACF and PACF (figure 4.7) shows that we cannot reject the hypothesis that residual ACFs are white noise.

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	8.76244	0.2093	41.8582	<.0001
Autoregressive, Lag 1	0.28472	0.0376	7.5634	<.0001
COS1	0.55129	0.3025	1.8222	0.0757
sin1	0.83728	0.2888	2.8991	0.0060
COS3	-0.48847	0.2921	-1.6721	0.1021
COS18	-0.20638	0.2859	-0.7219	0.4745
sin3	0.49135	0.2956	1.6624	0.1041
sin18	-0.08692	0.2859	-0.3040	0.7627
COS20	-0.09685	0.2847	-0.3402	0.7354
sin20	-0.61253	0.2850	-2.1491	0.0376
COS22	-0.15635	0.2827	-0.5531	0.5832
sin22	-0.14925	0.2822	-0.5289	0.5997
COS39	0.55539	0.2631	2.1112	0.0409
sin39	0.17463	0.2632	0.6635	0.5107
COS41	-0.48751	0.2606	-1.8708	0.0685
sin41	0.14719	0.2605	0.5649	0.5752
COS48	0.42532	0.2508	1.6962	0.0974
COS71	-0.21611	0.2222	-0.9726	0.3364
sin48	-0.56183	0.2509	-2.2390	0.0306
sin71	-0.07466	0.2220	-0.3363	0.7384
Model Variance (sigma squared)	14.63098	.	.	.

Figure 4.8 Parameter Estimation for Periodogram Model with AR(1) error model

Estimated parameters are shown in figure 4.8. The p-values of some coefficients are larger than 0.05, meaning that not all the coefficients are significant. Moreover, from figure 4.9-4.11, we find that there is periodicity in independent variables like

Temp, Pres and Dewp, which can be easily predicted by weather bureau and can also be easily obtained. Therefore, we would like to fit the series with the independent variables and periodogram.

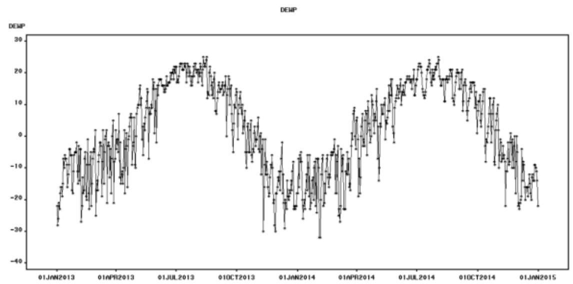

Figure 4.9 Plot of Dewp

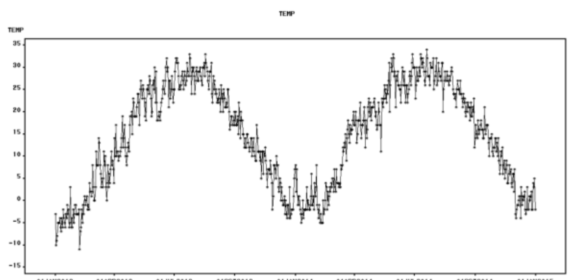

Figure 4.10 Plot of Temp

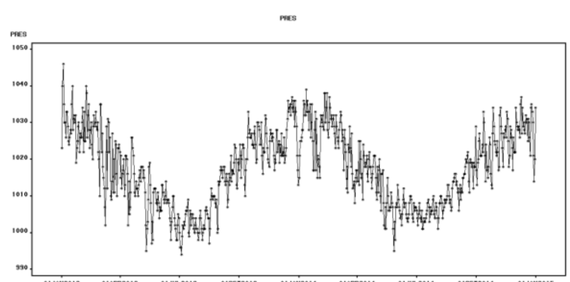

Figure 4.11 Plot of Pres

4.3 Building models for periodogram and independent variables

The result of ACF and PACF (figure 4.13) shows that we cannot reject the hypothesis that residual ACFs are white noise.

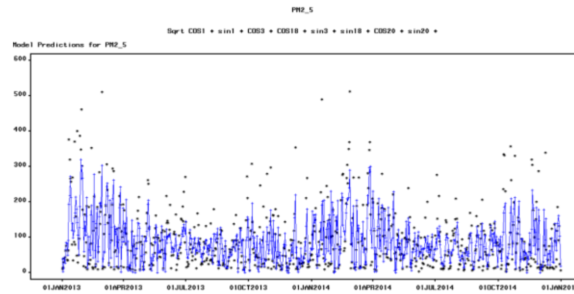

Figure 4.12 Actual versus fitted values of Periodogram Model and Independent variables with AR(1) error model

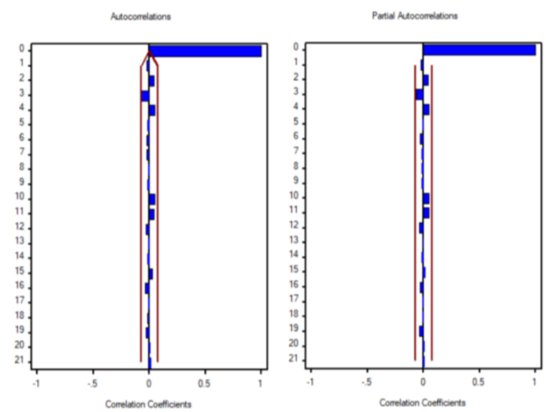

Figure 4.13 ACF and PACF of the residuals for Periodogram Model and Independent variables with AR(1) error model

Estimated parameters are shown in figure 4.14. After including Independent variables, almost all p-values of coefficients of periodogram model are larger than 0.05. On the other hand, all of the independent variables are significant. Therefore, it suggests that the independent variables can replace and even perform better than the Periodogram Models. Therefore, we would like to fit the series with only these independent variables.

Model Parameter	Estimate	Std. Error	T	Prob>¦T¦
Intercept	57.85997	23.4118	2.4714	0.0182
Autoregressive, Lag 1	0.26089	0.0381	6.8396	<.0001
COS1	9.28214	0.3973	23.3644	<.0001
sin1	3.63901	0.2147	16.9484	<.0001
COS3	-0.25803	0.1895	-1.3616	0.1816
COS18	-0.08698	0.1861	-0.4674	0.6430
sin3	-0.01015	0.1971	-0.0515	0.9592
sin18	-0.02620	0.1862	-0.1407	0.8889
COS20	-0.02844	0.1856	-0.1532	0.8790
sin20	-0.31030	0.1861	-1.6677	0.1038
COS22	-0.01353	0.1848	-0.0732	0.9420
sin22	-0.10123	0.1840	-0.5501	0.5855
COS39	0.27698	0.1745	1.5874	0.1209
sin39	-0.03911	0.1732	-0.2257	0.8226
COS41	-0.13527	0.1720	-0.7866	0.4365
sin41	0.30854	0.1737	1.7762	0.0839
COS48	0.02997	0.1664	0.1801	0.8581
COS71	-0.22017	0.1487	-1.4811	0.1471
sin48	-0.25132	0.1663	-1.5112	0.1392
sin71	-0.00140	0.1488	-0.009439	0.9925
DEWP	0.47418	0.0185	25.6677	<.0001
Is	-0.65500	0.2253	-2.9067	0.0061
Ir	-0.24682	0.0828	-2.9823	0.0050
PRES	-0.04896	0.0230	-2.1289	0.0400
Model Variance (sigma squared)	6.55688	.	.	.

Figure 4.14 Estimated model for Periodogram Model and Independent variables with AR(1) error model

4.4 Models with independent variables

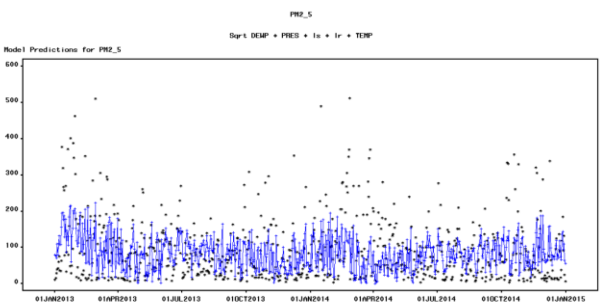

Figure 4.15 Actual versus fitted values of model of independent variables

The series plot with fitted values (figure 4.15) doesn't show a good fit either, since it cannot capture the high variance.

Obviously, residual for model of independent variables is not white noise (figure 4.16). From the ACF and PACF results, we would suggest to have an error model on AR(1).

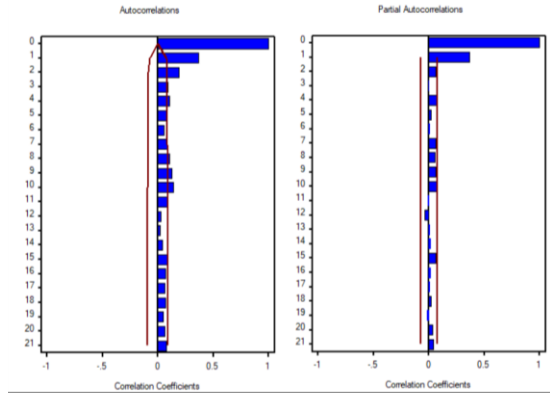

Figure 4.16 ACF and PACF of Model of Independent variables

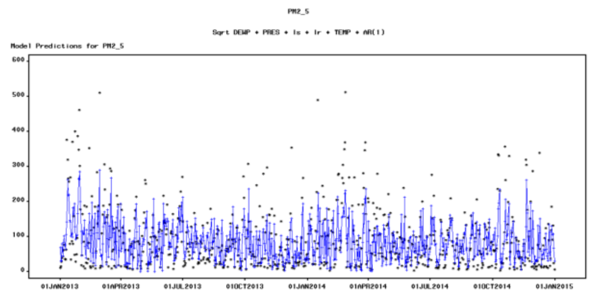

Figure 4.17 Actual versus fitted values of Model of Independent variables with AR(1) error model

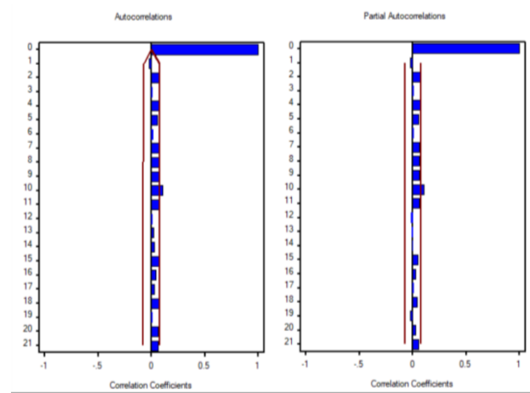

Figure 4.18 ACF and PACF of the residuals for model of independent variables with AR(1) error model

The series plot with fitted values (figure 4.17) show a better fit than the one without error model. The ACF, PACF of residual (figure 4.18) indicates that we cannot reject the hypothesis that residual is white noise.

Model Parameter	Estimate	Std. Error	T	Prob>¦T¦
Intercept	125.33386	27.2933	4.5921	<.0001
Autoregressive, Lag 1	0.43560	0.0355	12.2642	<.0001
DEWP	0.33886	0.0187	18.1417	<.0001
PRES	-0.10843	0.0266	-4.0692	0.0002
Is	-0.59510	0.2408	-2.4717	0.0166
Ir	-0.44547	0.0905	-4.9205	<.0001
TEMP	-0.46416	0.0257	-18.0932	<.0001
Model Variance (sigma squared)	8.52296	.	.	.

Figure 4.19 Estimated model for Model of Independent variables with AR(1) error model

The p-values of all coefficients are less than 0.05 (figure 4.19), meaning that all the coefficients are significant.

4.5 Model comparison

Model	Hold-out (RMSE)	Period of fit (RMSE)
Periodogram model	81.021	3.98
Periodogram model with AR(1) error model	78.637	3.83
Periodogram model with independent variables and AR(1) error model	58.403	2.56
Model of independent variables	54.185	3.20
Model of independent variables with AR(1) error model	52.305	2.92

Table 4.1 RMSE of hold-out samples and period of fit of models

Based the RMSE of hold-out samples, the model of independent variables with AR(1) error model performs the best, while Periodogram model with independent variables and AR(1) error model does the best in period of fit.

5. Comparison for all models

Model	Hold-out (RMSE)	Period of Fit (RMSE)
Seasonal dummies	81.911	4.05
Seasonal dummies with error model	79.352	3.86
Simple Exponential Smoothing	82.239	4.10
AR(1)	78.423	3.89
TF model on PRES	74.426	3.52
TF model on Iws	78.928	3.78
TF model on PRES and Iws	73.045	3.45
Periodogram model	81.021	3.98
Periodogram model with AR(1) error model	78.637	3.83
Periodogram model with independent variables and AR(1) error model	58.403	2.56
Model of independent variables	54.185	3.20
Model of independent variables with AR(1) error model	52.305	2.92

Table 5.1 Comparison for all models

Table 5.1 shows the root mean square error based on both hold-out and fit periods. We can conclude that model with independent variables and AR(1) error performs the best among all models.

Although including many independent variables in a time series model may bring in more uncertainty, we think it is a problem that can be overcame in our case, since the meteorological variables can be easily predicted by weather bureau and can also be easily obtained with highly mature techniques and theories. Besides, instead of using periodogram model, the model with independent weather variables is much easier to interpret, since the level of air pollution is highly related to weather.

There are two possible next steps for our project. First is to verify the prediction accuracy of our model after the data of 2015 is published. Second is to use hourly data to catch more changes in PM2.5, and build better models in prediction.