DNSC 6219 SPRING 2019

Section 10

Time Series Analysis of Particulate Matter (PM2.5) Trends

Group member: Xu Jiaru, Wang Xinyu, Wu Jingsi, Liu Yuqi, Li Weiteng

1. Introduction and Overview

With the development of the economy, living standard has been improved at the same time. People tend to pay more attention to the environment around us instead of on making more money. However, five years ago in China, as the regulation was not as complete as it is now, a lot of companies grow as a cost of environment which greatly contribute to the decline in air quality over the recent years. In particular, in Beijing, China, haze pollution occurs frequently during particular seasons and it raises continuing public concerns over air quality. Baes on all previous concerns, we decided to probe deep into the air quality pattern in China over the previous three years to see whether the pollution worsens over the years or has already been under control.

To conduct our analysis, we collect daily observation data from University of California, Irvine Machine Learning Repository. For PM2.5 concentration observations, they are originally gathered by U.S. Department of State. The dataset also includes several other variables such as daily temperature, humidity and wind speed which may help us to gain more insights into the air pollution situation. With 1095 observations in total, we set our hold out level to 100 periods. In this mini project, we will use knowledge from Time Series Forecasting class to analyze and predict the series. Hopefully what we do can make the world better.

As for our dataset, we decided to use one that describes the air quality of Beijing from year of 2013 to 2015. The criteria that we used to measure air quality is the level of PM2.5, which represents the average atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers. A lower PM2.5 level refers to a better air quality and a higher PM2.5 denotes a more severe air pollution. In our dataset, the main series we are interested in modeling is the variable called *PM_US_POST*, which indicates the PM2.5 concentration level.

In this section, let's first consider the daily data on "PM2.5 concentration" in Beijing, starting from January 1st, 2013, and start our analysis with the time series plot and the seasonal box plot by both month and week.

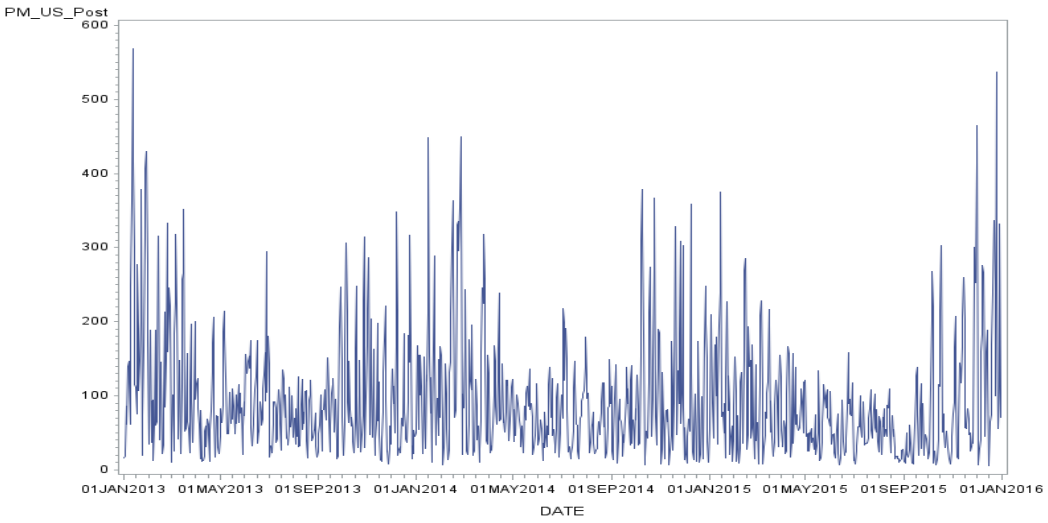

Figure 1.1: Overall Time Series Plot

From Figure 1.1, 1.2, 1.3, we can have a straightforward understanding into our dataset. In Figure 1.1, although there is not a distinct trend over the three years, we clearly observe some yearly iteration over the three years. At the beginning and the end of each year, there is always more variation within the air pollution and more extreme values appear. However, a lower level of air pollution always takes place in the middle of each year with less variation.

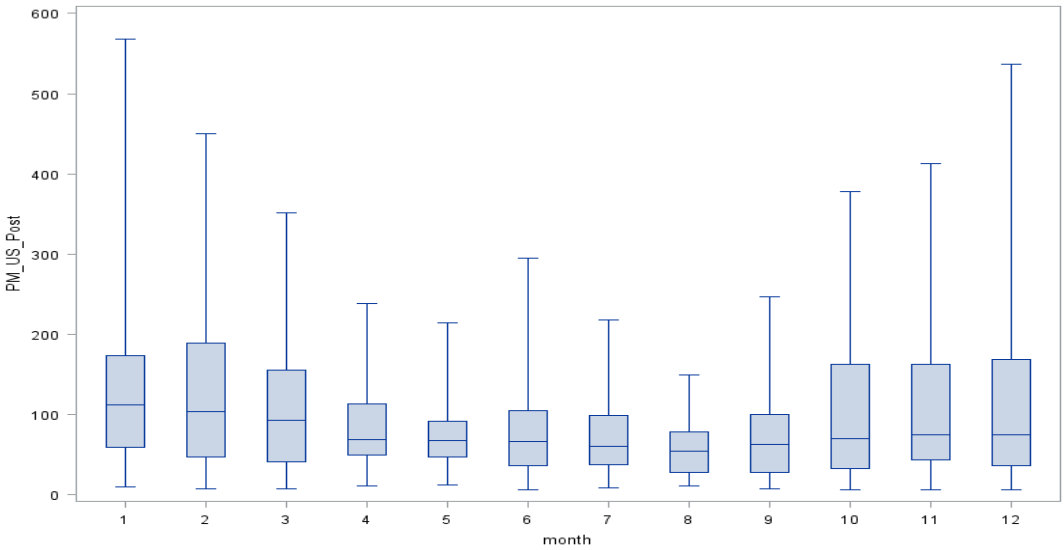

Figure 1.2: Monthly Seasonal Boxplot

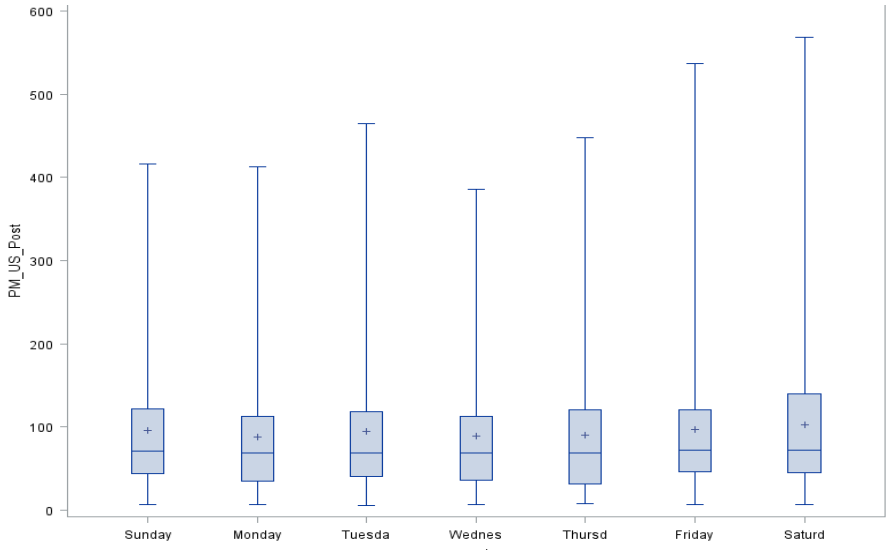

Figure 1.3: Weekly Seasonal Boxplot

When we explore into the seasonality pattern with Figure 1.2 and Figure 1.3, we find out that the monthly plot reveals more information than the weekly boxplot. Figure 1.2 describes our intuition from the overall picture more detailly. More specifically, from October to March, the overall PM2.5 level is slightly higher than the PM2.5 level from April to September. According to the box plot, we can easily observe that the range of PM2.5 level is significantly different for each month. For January and December, the highest PM2.5 level can reach up to the level of 500+, yet the highest PM2.5 level for August is not even close to the level of 200. From Figure 1.3, we can conclude that weekdays may not have a much more significant influence on air quality than the distribution of month.

2. Univariate Time-series models

2.1 Deterministic Time Series Models and Error model.

From the analysis above, we can conclude that there is some seasonality within our time series. To fit a model that can reflect seasonal variation, we have multiple choices and various kinds of models to choose from. To begin with, we will first fit our series with seasonal dummies and trend model and then check for the errors.

2.1.1 Deterministic Time Series Model

For deterministic time series models, we can fit either seasonal dummy variables or trigonometric functions to fit the series. First, let's consider about the seasonal dummy model.

In order to construct our seasonal dummy model, we first build monthly dummies and another set of dummy variables that indicates whether it is winter season or not according to our previous observation through monthly boxplot. Then, from the parameter estimates in Table 2.1.1, we can conclude that the majority of parameters are significant. Linear trend is significant with an extremely small negative value which denotes that pm concentration is slowly decreasing in the long term.

PM_US_POST

Seasonal Dummy model

Model Parameter	Estimate	Std. Error	T	Prob > |T|
Intercept	70.21405	10.1153	6.9414	<.0001
JAN	42.25346	11.8573	3.5635	0.0006
FEB	34.84803	12.0838	2.8839	0.0050
MAR	95.61205	11.8168	8.0912	<.0001
APR	69.34216	11.8818	5.8360	<.0001
MAY	63.24319	11.7955	5.3616	<.0001
JUN	66.72566	11.8711	5.6209	<.0001
JUL	63.94646	11.7952	5.4214	<.0001
AUG	50.67692	11.8030	4.2935	<.0001
SEP	66.09653	12.1131	5.4566	<.0001
OCT	113.16806	12.9279	8.7538	<.0001
NOV	88.53611	13.0279	6.7959	<.0001
DEC	-2.87993	0	.	
Winter	42.04652	0	.	
nowinter	-42.04652	0	.	
Linear Trend	-0.03756	0.0083	-4.6035	<.0001
Model Variance (sigma squared)	5173	.	.	.

Table 2.1.1: Parameter Estimates with Seasonal Dummy model

Second, about Cyclical Trend model, recalling our analysis into the overall series in Section 1, there still seems exist some seasonality other than by week. Therefore, we also attempt to fit our model with trigonometric functions.

Obs	FREQ	PERIOD	P_01
4	0.01721	365	581002.55
6	0.02869	219	133157.46
68	0.38445	16.34	121564.12
78	0.44183	14.22	95752.81
124	0.70578	8.9	93205.04
144	0.82054	7.66	90219.3
3	0.01148	547.5	84971.01
145	0.82628	7.6	82661.63
127	0.723	8.69	63930.53
114	0.6484	9.69	63774.39
29	0.16067	39.11	63042.94
32	0.17788	35.32	62152.14
54	0.30412	20.66	62103.19
147	0.83776	7.5	61966.58
188	1.07302	5.86	60499.38

Table 2.1.2: Top 15 Periodograms and associated Periods

Table 2.1.2 declares the top 15 periodograms and their associated periods in our model. For further exploration, we pick harmonics with top six periodogram values, respectively 3,5,67,77,123,143.

After fitting cyclical trend model on our series, we observe that it provides a forecast with a clear drop even down to negative values and fits our data with cyclical trends over time. However, it does not reflect the information well for values that are high or around zero. The model mainly modifies the PM2.5 level in the middle area.

Bearing an understanding of the general predictability of cyclical trend model, we desire to check the parameter estimates as shown in Table 2.1.3. From the parameter table, we can conclude that there is still a significant downward trend within the series. From the model performance, compared with the seasonal dummy model, we can see that cyclical trend model has a lower model variance. Therefore, cyclical trend model provides a better fit for our data then the seasonal dummy one.

PM_US_POST

Cyclical Trend model

Model Parameter	Estimate	Std. Error	T	Prob > |T|
Intercept	113.84354	4.5659	24.9334	<.0001
COS3	27.76321	3.2217	8.6175	<.0001
SIN3	6.23987	3.2445	1.9232	0.0578
COS5	12.41886	3.1554	3.9357	0.0002
SIN5	0.36202	3.2229	0.1123	0.9108
COS67	-10.44625	3.1617	-3.3040	0.0014
SIN67	-11.06811	3.1567	-3.5062	0.0007
COS77	-2.17519	3.1597	-0.6884	0.4930
SIN77	-9.76543	3.1587	-3.0916	0.0027
COS123	-7.93293	3.1608	-2.5098	0.0140
SIN123	5.66392	3.1573	1.7939	0.0763
COS143	-12.04772	3.1597	-3.8130	0.0003
SIN143	-6.85625	3.1584	-2.1709	0.0327
Linear Trend	-0.04116	0.0080	-5.1731	<.0001
Model Variance (sigma squared)	4964	.	.	.

Table 2.1.3: Parameter Estimates with Cyclical Trend model

2.1.2 Error Model

As we previously fit both seasonal dummy model and cyclical trend model to our series, we find that seasonal dummy model has a RMSE of 118.23 while that of cyclical trend model is 104.6644. Besides, cyclical trend model also has a lower model variance which denotes for a

better fit over the training set. Therefore, we can conclude that cyclical trend model provides a better fit than seasonal dummy model. Next, we are going to take a deeper look into cyclical trend model.

Prediction error autocorrelation plot for cyclical trend model is shown in Figure 2.1.1. By simply looking at the ACF and PACF plot, we can see the ACF of PM2.5 level is fast-decaying, so we could say that the series is stationary. Then as the PACF of PM2.5 level truncated after the second lag, we might consider the AR (2) or ARMA(1,1) model for PM2.5 concentration level. Also, observing the partial autocorrelation at lag 2 has already been relatively small, AR (1) model might also be considerable for fitting our errors under such circumstances.

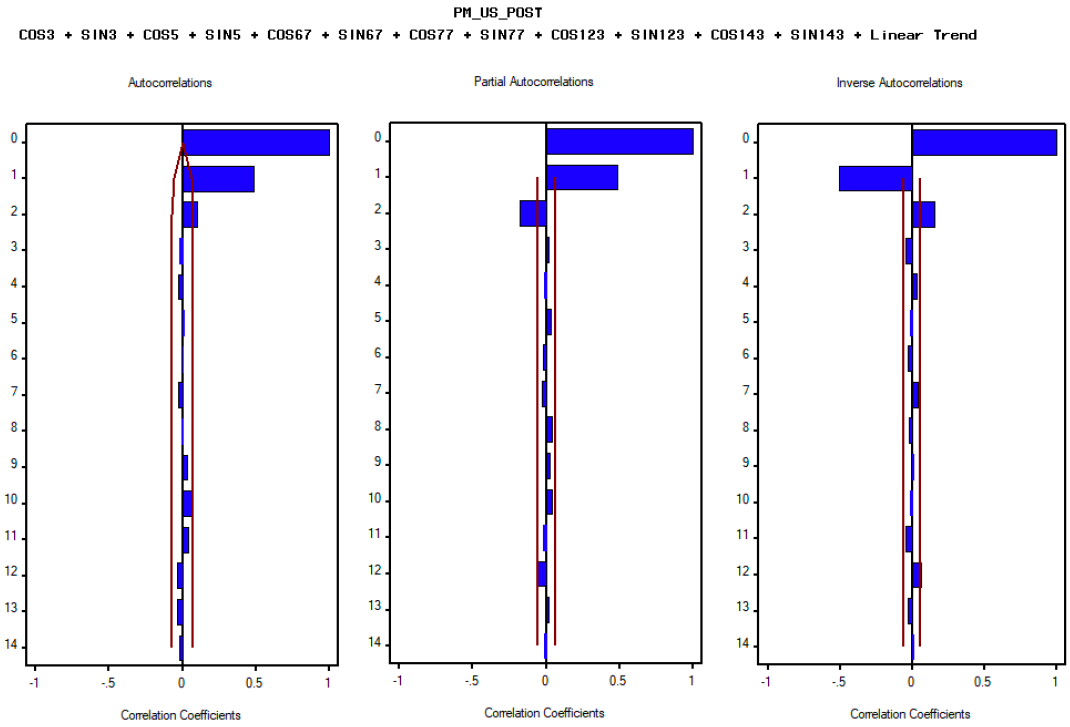

Figure 2.1.1: Prediction Error Autocorrelation Plots of Cyclical Trend model

After we fit different models to our error terms, we observe that ARMA(1,1) provides a better fit on our hold out sample than the other two. Figure 2.1.2 provides the actual versus predicted plots for the model. From the plot, we can clearly see that the model provides a prediction with more fluctuations and is closer to the original observations. Other than only fetch the information with a certain range, the ARMA error terms help the model to provide a better fit to our data with

more variations over the three years. Then checking for the behavior of residuals, we can be reasonably sure that the residuals are white noise here as expected through the ACF plot shown in Figure 2.1.3.

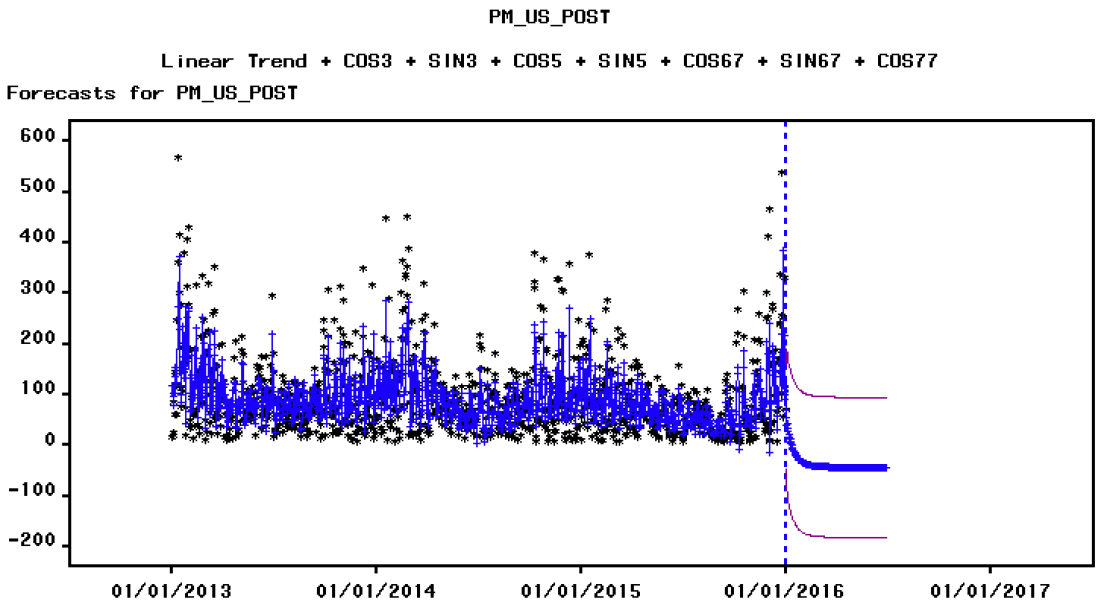

Figure 2.1.2: Actual versus Predicted Plot of Error model

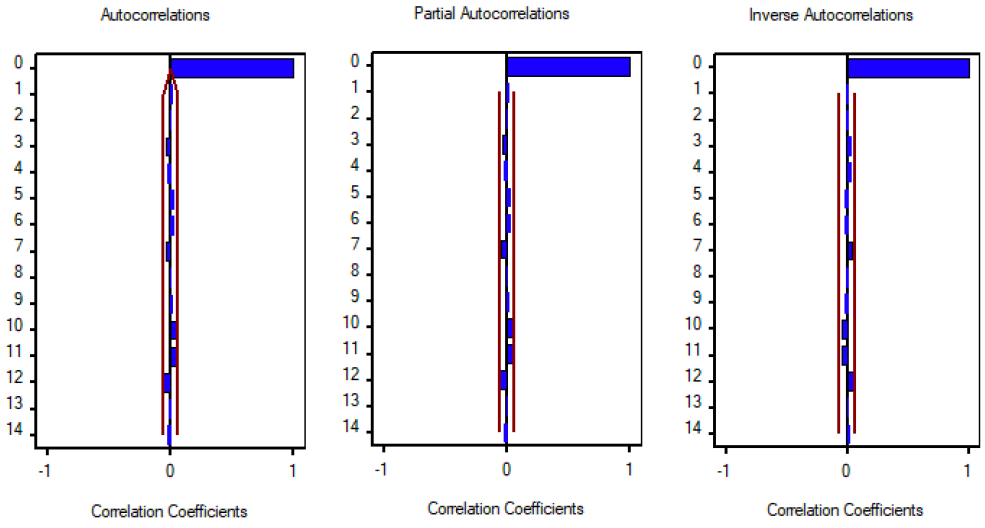

Figure 2.1.3: Prediction Error Autocorrelation Plots with Error model

Through the previous process, we fit our time series with a cyclical trend model and fitting the residuals with an ARMA(1,1) model and after which, we have successfully reduced the residuals to white noise. The detailed parameter estimates are shown in Table 2.1.4.

PM_US_POST

Cyclical Trend Error model

Model Parameter	Estimate	Std. Error	T	Prob > |T|
Intercept	113.53192	6.9111	16.4275	<.0001
Autoregressive, Lag 1	-0.32929	0.0569	-5.7860	<.0001
Autoregressive, Lag 2	0.24417	0.0584	4.1787	<.0001
Linear Trend	-0.04078	0.0120	-3.3663	0.0012
COS3	27.66466	4.8789	5.6702	<.0001
SIN3	6.15593	4.9132	1.2529	0.2137
COS5	12.19208	4.7753	2.5531	0.0125
SIN5	0.32831	4.8818	0.0673	0.9465
COS67	-10.44974	4.5802	-2.2815	0.0250
SIN67	-11.07657	4.5772	-2.4200	0.0177
COS77	-2.18449	4.5151	-0.4838	0.6298
SIN77	-9.70380	4.5184	-2.1476	0.0346
COS123	-7.94852	4.1635	-1.9092	0.0597
SIN123	5.62959	4.1604	1.3531	0.1796
COS143	-12.06943	3.9824	-3.0307	0.0032
SIN143	-6.76781	3.9846	-1.6985	0.0931
Model Variance (sigma squared)	3687	.	.	.

Table 2.1.4: Parameter Estimates with Error model

2.2 ARIMA models

From the plot of our original series, we can clearly discover that although observations are densely distributed around 0 to 100, there still remains some monthly variations along the three years. Recalling the autocorrelation, partial autocorrelation and inverse autocorrelation plots, we notice that the partial autocorrelation is clearly chopping off after lag 1 but sill beyond the 2

standard deviation bound with lag 2, the inverse autocorrelation behaves similarly and as for the autocorrelations, we can take its behavior as decaying exponentially to some extent. Therefore, we can fit AR (2) model first and evaluate its performance and the residual pattern.

2.2.1 Non-seasonal ARIMA

From the Actual versus Predicted plots as in Figure 2.2.1 below, we notice that the model provides more accurate predictions for fine particles ($PM_{2.5}$) at lower level than at higher level when it is over 300ug/m^3. In another words, although autoregressive model can make adjusted predictions based on the previous observations, it still cannot learn "extreme values" within the series. For the prediction horizon, we notice that as we step further, the predictions will converge to mean and at smaller lags because the predictions and observations based on are still located within the same month, there is not much variation and fluctuations revealed.

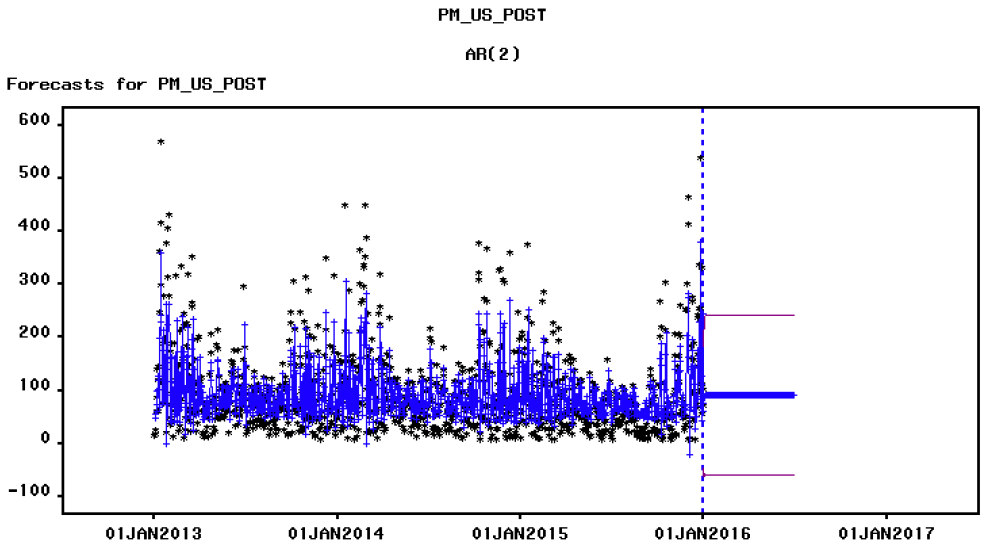

Figure 2.2.1: Actual versus Predicted Plot with AR (2) model

After we look into the actual versus predicted plot, let's transfer to the autocorrelation plots as shown in Figure 2.2.2. From the three plots, we observe that autocorrelation, partial autocorrelation and inverse autocorrelation stay within the 2 standard error bound since lag 1. Therefore, we are reasonably sure that residuals here are white noise.

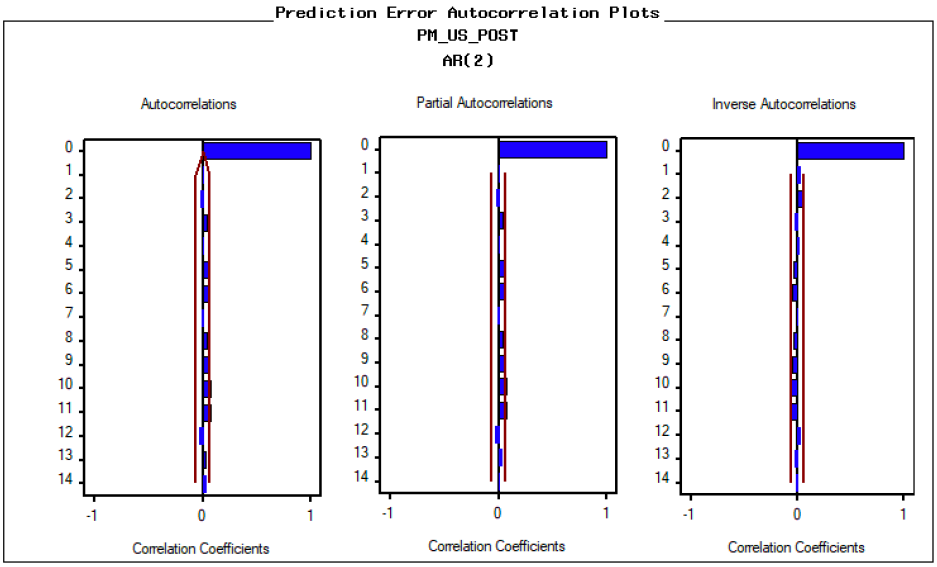

Figure 2.2.2: Prediction Error Autocorrelation Plots with AR (2) model

From the parameter estimates in Table 2.2.1, we observe that all the parameters are smaller than 0.05. Therefore, we can be 95% confident that those parameters are significantly different than zero in the population.

PM_US_POST

AR (2)

Model Parameter	Estimate	Std. Error	T	Prob > |T|
Intercept	92.05452	3.9327	34.4072	<.0001
Autoregressive, Lag 1	0.65087	0.0314	20.7580	<.0001
Autoregressive, Lag 2	-0.15673	0.0314	-4.9989	<.0001
Model Variance (sigma squared)	3945	.	.	.

Table 2.2.1: Parameter Estimates with AR (2) model

2.3 Comparison of models

With all types of deterministic and ARIMA models as discussed above, we wonder which provides the best fit towards our data. Here for comparison, we use the Root Mean Square Error as the criteria. The detailed RMSE for each model is showed in Table 2.3.1 below.

Model	Model Title	RMSE	Model Variance
1	AR (2)	92.13918	3945
2	Cyclical Trend Model with AR (2) Error	89.71496	3668
3	Cyclical Trend Model with ARMA (1,1) Error	88.78375	3687
4	Cyclical Trend Model	104.66444	4964
5	Seasonal Dummy Model	118.23008	5173

Table 2.3.1: Comparisons on Model Performance

From Table 2.3.1, we discover that when without error model, cyclical trend model provides a better fit towards both training set and hold out samples. For error term, ARMA(1,1) offers a slightly better performance on hold out sample.

In conclusion, when we assign more focus to our hold out sample, we will have cyclical trend model with an ARMA(1,1) error the best model for our series with an RMSE of 88.78375 among all kinds of models that we build under a hundred hold-out samples.

3. Multivariate Time Series Models

First, for multivariate model, we decide to pick up variables among temperature, humidity and wind speed as potential predictors and they are named as TEMP for temperature, HUMI for humidity and Iws for wind speed in our series.

3.1 Model Identification

In order to set up our multivariate time series model, we have to go through several steps to make sure that they are qualified to perform as a regressor in our model. In the following part, we are going to go through each variable in the following sequence: temperature humidity and

windspeed, check for stationarity and if the input series is white-noise and then identify the model according to the crosscorrelation.

3.1.1 Temperature

In our series, temperature represents the overall air temperature in a certain day measured by Celsius degree. Intuitively, according to our common sense, temperature is correlated with each other and today's temperature is affected by yesterday's temperature and follow a cycle of 365 days. Plot for temperature over three years is shown in Figure 3.1.1. From the three bell-shape in the plot, we notice that temperature follows a yearly period. When we look into the temperature series in our dataset, we notice that its autocorrelation is decaying slowly with the lags as in Figure 3.1.2 and therefore, we take the first difference, after which the series becomes stationary as in Figure 3.1.3.

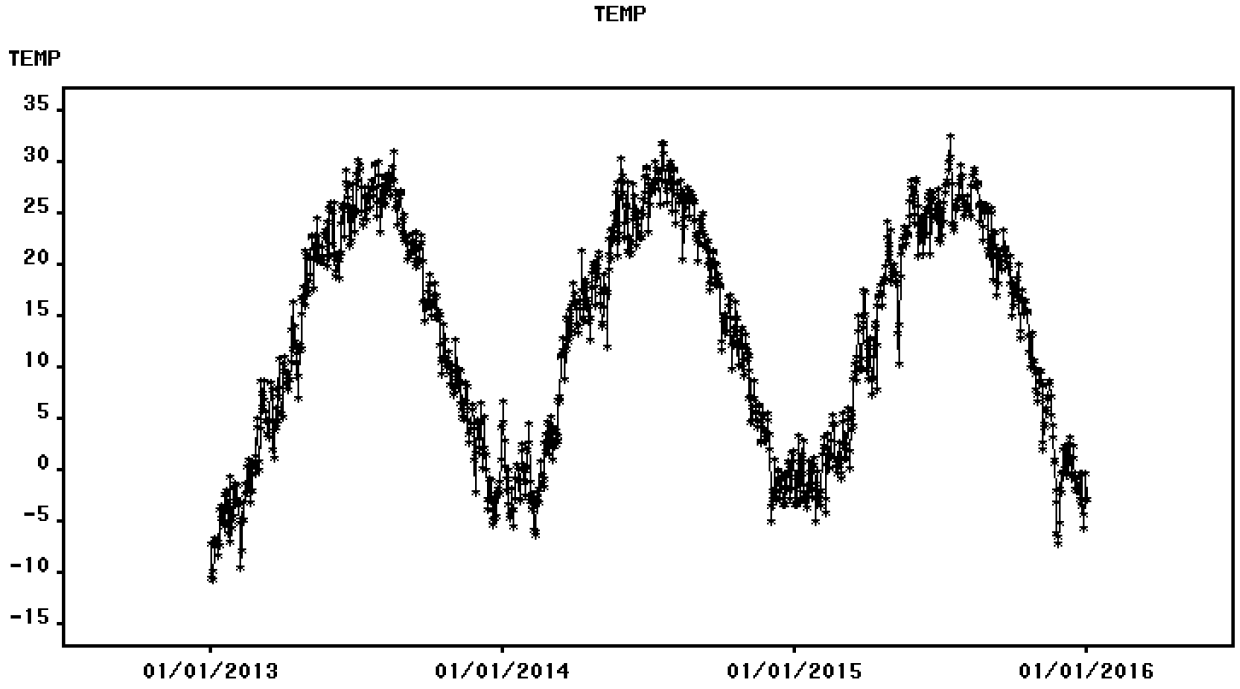

Figure 3.1.1: Temperature Plot

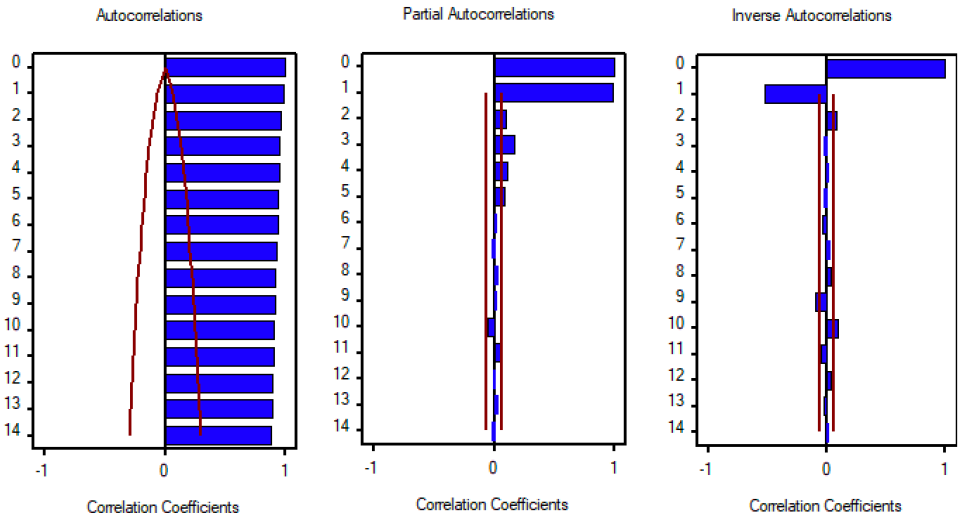

Figure 3.1.2: Prediction Error Autocorrelation Plots with Temperature

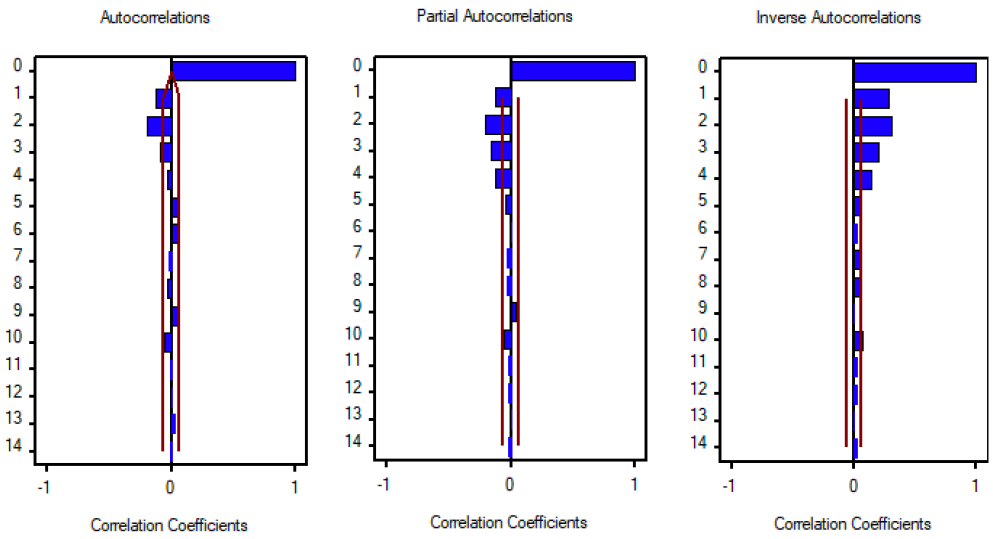

Figure 3.1.3: Prediction Error Autocorrelation Plots with first difference Temperature

Then, for the next step, we check if the series is already white noise. Observing that both its ACF and PACF show as decay, we fit an ARMA (2,2) model and make it a white noise successfully. After making sure that temperature series is a white noise, we can take the crosscorrelation function proportionally to impulse response weight and then to further estimate the parameters.

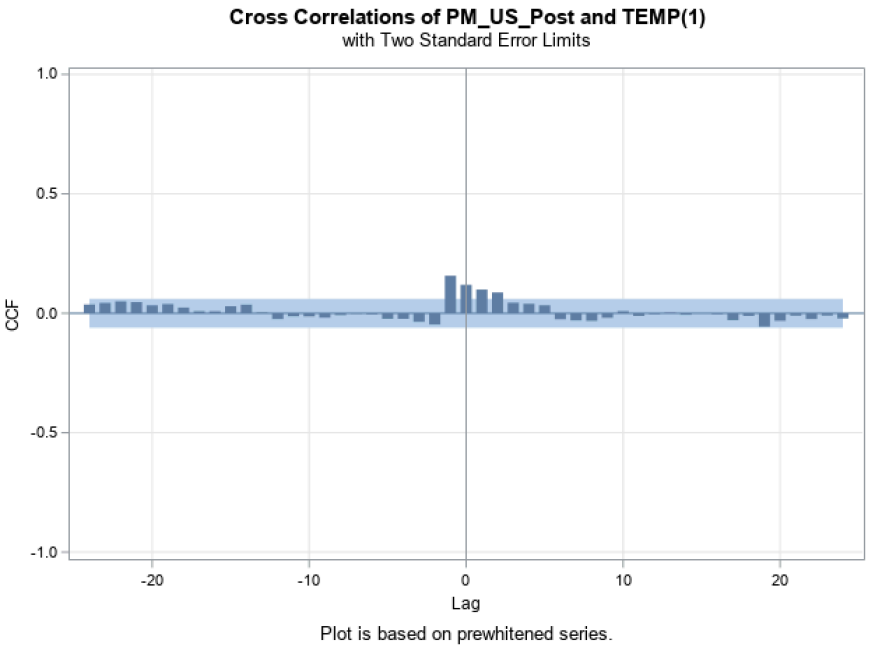

Figure 3.1.4: Cross Correlation Plot of PM2.5 and Temperature

From Figure 3.1.4, we can see that the cross correlation between those two variables are most significant at negative lag and then slowly decay to positive lags. Therefore, with negative significance, we cannot parallel cross correlation functions to impulse response weights. The scenario can be explained that there exists a reversed causality between air pollution and temperature. It is not simply the situation where temperature can explain air quality but on the other side, air pollution could also be one potential cause for temperature increase with some biological reasons behind.

3.1.2 Humidity

Humidity here is another variable included in our dataset and it measures the moisture in the air by percentage. Following similar reasons as temperature, humidity is also correlated between days and not stationary. Therefore, we take the first difference of the original series and then transform it into a stationary series. Detailed trend and correlation analysis plots are shown below in Figure 3.1.5.

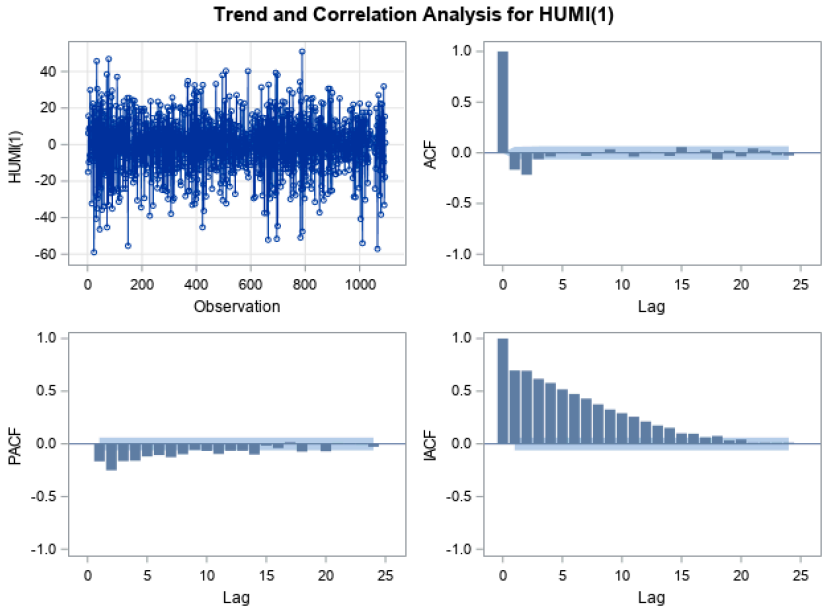

Figure 3.1.5: Trend and Correlation Analysis for simple difference of Humidity

From the correlation plots, we can clearly identify a decaying inverse autocorrelation plot and a chopping off autocorrelation plot and therefore, fit an autoregressive model at order 2 to reduce the humidity series to white noise. After performing such transformation, we get the crosscorrelation plot as shown in Figure 3.1.6. Similarly, we observe cross correlation plot significant at several negative lags and therefore, humidity does not qualify for a transfer function model.

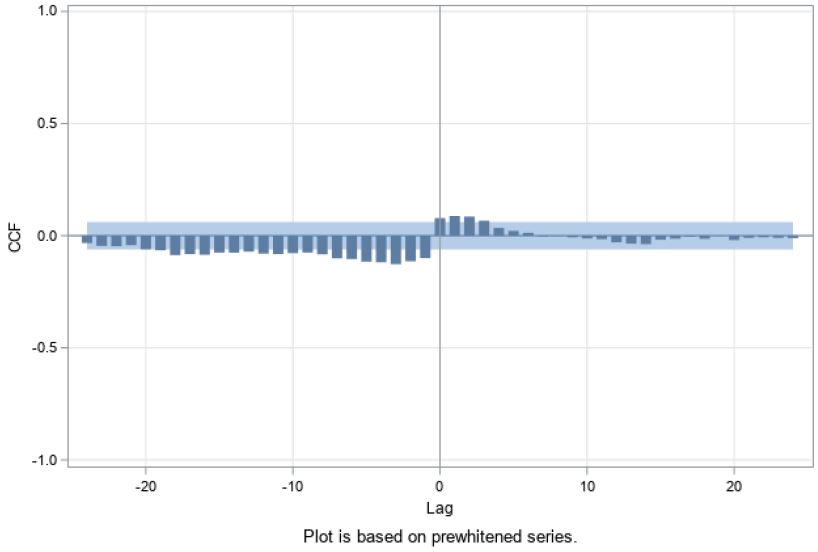

Figure 3.1.6: Cross Correlation Plot of PM2.5 and Humidity

3.1.3 Wind Speed

In our dataset, we have cumulated wind speed as one of our variables and it is measured in meters per second. Intuitively, with a higher wind speed, pollution particles scattered in the air will be more likely to be blown off which, as a result, decreases the concentration of PM2.5 particles in the air and improves air quality.

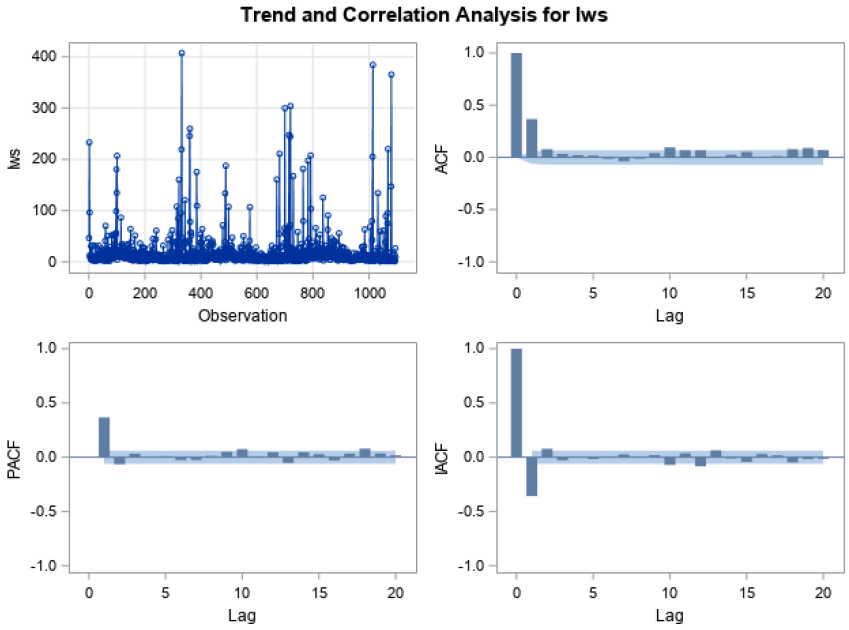

Figure 3.1.7: Trend and Correlation Analysis for Wind Speed

Figure 3.1.7 provides us with a discrete insight into the wind speed series with trend and correlation analysis. From the autocorrelation plot, we can observe that it can be interpreted as a decaying behavior and from the partial autocorrelation plot and the inverse autocorrelation plot, we can conclude that partial autocorrelation is chopping off after lag 1. Therefore, we decide to fit an AR (1) model trying to pre-whiten the wind speed series. With autocorrelation check for residuals as in Table 3.1.1, we can be reasonably sure that the residuals are white noise which also denotes what we can take cross correlations to briefly estimate impulse response weights.

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	6.6	5	0.2525	0.024	0.07	0.007	0.015	0.015	-0.008
12	22.72	11	0.0194	-0.032	-0.008	0.016	0.087	0.018	0.074
18	32	17	0.015	-0.031	0.019	0.051	-0.015	-0.003	0.064
24	39.29	23	0.0184	0.054	0.051	0.009	-0.018	-0.024	0.001
30	51.49	29	0.0062	-0.012	0.007	-0.007	0.056	0.08	0.032
36	52.87	35	0.0268	-0.009	-0.014	-0.01	0.027	-0.001	-0.01
42	61.36	41	0.0213	-0.033	0.012	0.076	0.007	-0.021	0.004
48	65.88	47	0.0359	-0.033	0.018	-0.003	0.038	-0.019	0.026

Table 3.1.1: Autocorrelation Check of Residuals

After the above process, we have already reduced the input series to white noise. Then we tend to focus on the performance of cross correlation functions as shown in Figure 3.1.8. From this plot, we see that the first significant CCF comes at lag 0 and its behavior afterwards could be interpreted as decaying slowly. Therefore, we can take b=0, s=0 and r=1 as parameters in our transfer model.

Figure 3.1.8: Cross Correlation Plot of PM2.5 and Wind Speed

3.2 Model Estimation and Validation

In the previous part, we have successfully identified that cumulated wind speed could serve as an input variable in our transfer models with itself a white noise after some transformation. With parameters estimated from cross correlation functions, we can now build our transfer models. Then when we take a look at our fitted model, we notice that with a simple transfer model, the residuals are not white noise and through the autocorrelation and partial autocorrelation plots as in Figure 3.2.1, we identify an AR(2) process in an attempt to reduce the residuals to white noise and finally make our model a transfer noise model.

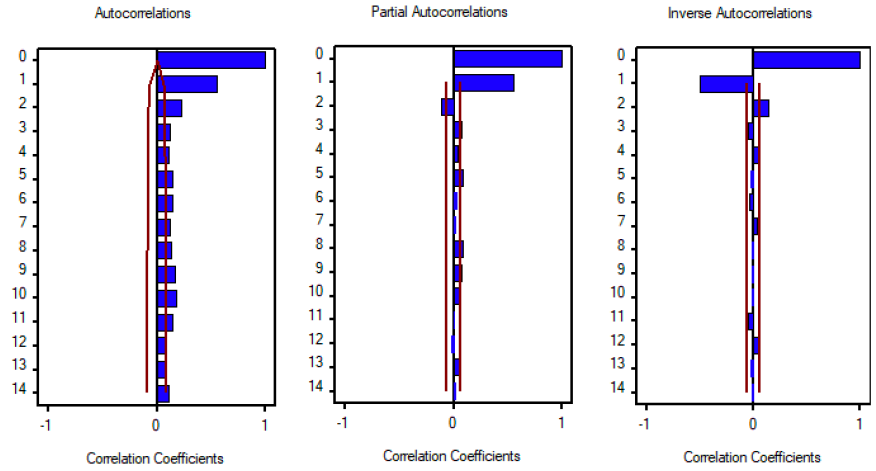

Figure 3.2.1: Prediction Error Autocorrelation Plots with TF model

After modifying for the residuals, we get the maximum likelihood estimation of parameters as shown in Table 3.2.1.

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	T Value	Approx. Pr >|t|	Lag	Variable	Shift
MU	108.96399	4.91213	22.18	<.0001	0	PM_US_POST	0
AR1,1	0.62287	0.03011	20.69	<.0001	1	PM_US_POST	0
AR1,2	-0.11647	0.03020	-3.86	0.0001	2	PM_US_POST	0
NUM1	-0.44037	0.05459	-8.07	<.0001	0	Iws	0
DEN1,1	0.39981	0.10520	3.80	0.0001	1	Iws	0

Table 3.2.1: Maximum likelihood Estimation of Parameters

After estimation of parameters, we also need to take into account the adequacy of transfer model and check for residuals and whether residuals are correlated with input variables.

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	15.68	4	0.0035	0.009	-0.034	0.033	0.015	0.079	0.074
12	38.91	10	<.0001	0.005	0.047	0.069	0.089	0.076	-0.017
18	55.21	16	<.0001	0.039	0.015	0.079	0.065	0.048	0.012
24	70.24	22	<.0001	0.053	-0.017	0.077	0.051	0.02	0.038
30	91.89	28	<.0001	0.049	0.05	0.003	0.086	0.052	0.065
36	105.42	34	<.0001	0.035	0.084	0.004	-0.027	0.021	0.05
42	116.28	40	<.0001	0.038	0.032	0.014	0.063	0.054	0.008
48	124.82	46	<.0001	-0.001	0.007	0.054	0.06	0.029	0.005

Table 3.2.2: Autocorrelation Check of Residuals

Crosscorrelation Check of Residuals with Input Iws									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
5	1.76	5	0.8806	0.001	-0.02	0.018	0.007	0.017	0.024
11	18.58	11	0.069	0.075	-0.025	0.01	0.071	0.061	-0.01
17	22.32	17	0.1726	0.001	0.058	-0.006	-0.006	0.002	-0.003
23	29.6	23	0.1612	0.036	-0.005	0.02	0.044	-0.051	0.019
29	43.06	29	0.045	0.02	0.068	0.04	0.072	-0.022	0.006
35	46.85	35	0.0869	0.019	-0.032	-0.004	0.034	0.011	0.029
41	49.05	41	0.1816	0.004	0.017	-0.005	-0.03	0.011	-0.025
47	51.75	47	0.2937	-0.008	-0.023	0.005	0.029	-0.03	-0.011

Table 3.2.3: Crosscorrelation Check of residuals with Input Iws

Table 3.2.2 and 3.2.3 together provide strong evidence over the properness of a transfer noise model. From the autocorrelation check of residuals, although the p-values are mostly significant, taking a second look into the detailed values of autocorrelations, we notice that the autocorrelations are not large, and the significance could be partially due to the number of observations we have. Under such consideration, we can reasonably conclude that the residuals are white noise. As for the correlation between residuals and input series, from the crosscorrelation check as in Table 3.2.3, we can tell that the input wind speed series is not

correlated with residuals. Therefore, we make an adequate application of transfer noise model here with wind speed series and PM2.5 series.

4. Conclusion

After our identification of transfer noise model, we now prefer to combine our regression model with several deterministic models in the previous section in order to fully improve the general performance of our model. Several model performances are showed in Table 4.1 as below.

Model	Model Title	RMSE	Model Variance
1	AR (2)	92.13918	3945
2	Seasonal Dummy Model	118.23008	5173
3	Seasonal Dummy with ARMA(1,1) and TF Model	86.76707	3525
4	Seasonal Dummy with AR(2) and TF Model	87.94336	3500
5	Seasonal Dummy Model with ARMA(1,1)	92.57457	3796
6	Cyclical Trend Model	104.66444	4964
7	Cyclical Trend with ARMA(1,1) Model	87.15351	3724
8	Cyclical Trend with ARMA(1,1) and TF Model	85.07846	3457
9	Cyclical Trend with AR(2) and TF Model	88.39528	3396

Table 4.1: Model Comparisons with RMSE and Model Variance

From the Table 4.1 above, we notice that cyclical trend with ARMA(1,1) and TF-noise model performs best on predicting our hold-out sample while cyclical trend with AR(2) and TF-noise model does a better job at fitting our training data. Besides, through comparisons over various models listed in the table, we can also conclude that adding regression is making our model performs better to different extent. For our seasonal dummy model, we can expect a more significant drop in RMSE while a minor drop on RMSE happens to our cyclical trend model.

Compared with univariate time-series model we fit in section 2, we have achieved a significant improvement on model performance. Next, let’s take a deeper look into our champion model selected based on RMSE.

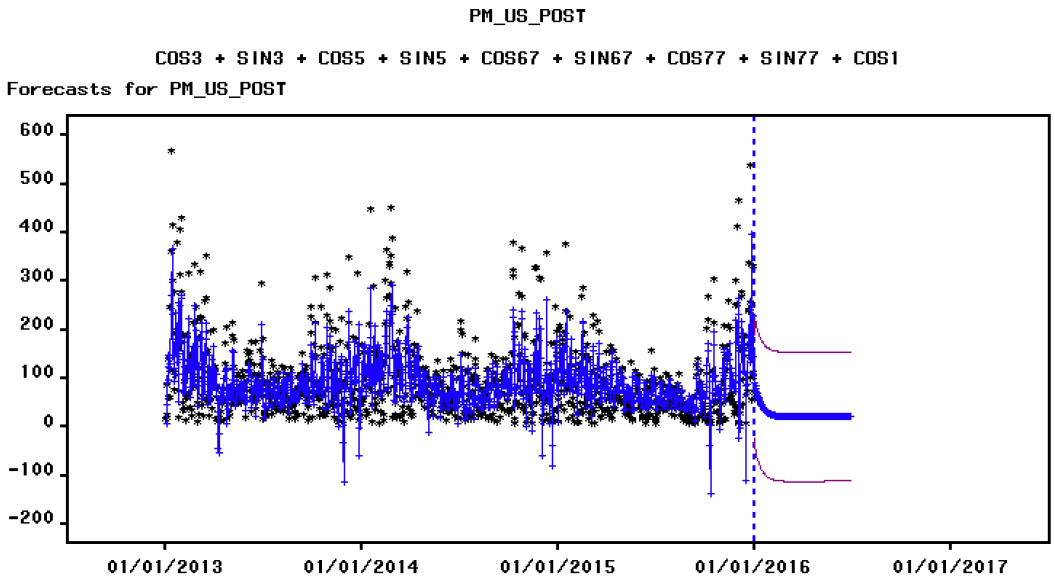

Figure 4.1: Actual versus Predicted Plot for our champion model

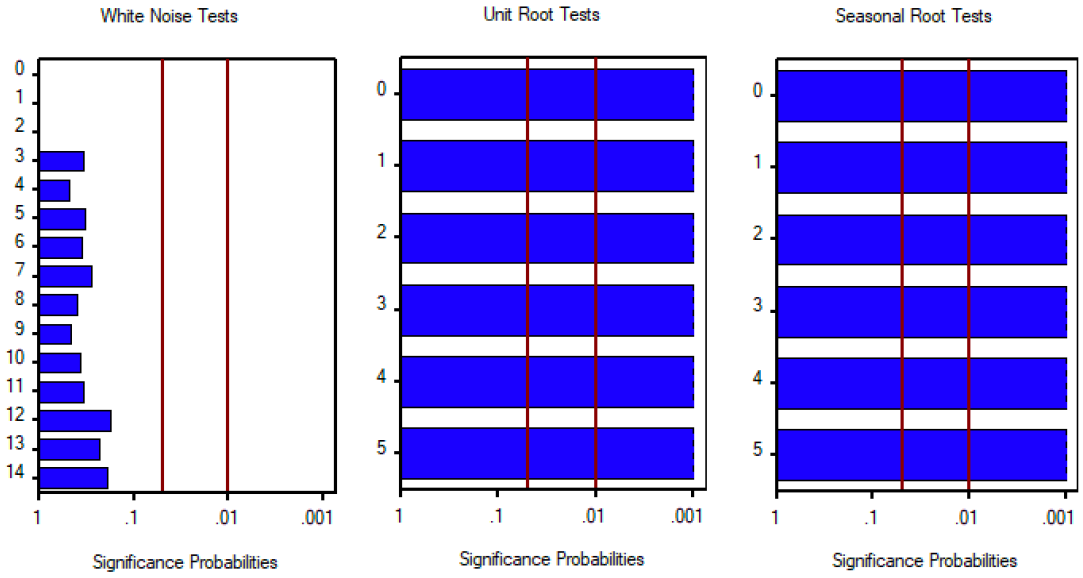

Figure 4.2: Prediction Error White Noise/Stationarity Test Probabilities

From the actual versus predicted plot in Figure 4.1, we notice that model does a preferred job for predicting for our hold out samples which is close to spring season. The model comes up with predictions lower than that for winters. Moreover, it does a better job at catching variations and extreme values. Although it fits some periods with negative values which is against our common sense, actual values associated with those negative predictions are mostly close to zero. The

autocorrelations and partial autocorrelations are both stationary and stay within the 2 standard error boundaries across all lags. From Figure 4.2, we further prove that the residuals are already white noise.

PM_US_POST				
Cyclical Trend with ARMA(1,1) and TF Model				
Model Parameter	Estimate	Std. Error	T	Prob > |T|
Intercept	113.11036	4.7040	24.0455	< .0001
Moving Average, Lag 1	-0.29678	0.0596	-4.9773	< .0001
Autoregressive, Lag 1	0.25612	0.0603	4.2481	< .0001
COS3	38.14746	4.8121	7.9274	< .0001
SIN3	11.33134	4.7042	2.4088	0.0182
COS5	9.19582	4.6343	1.9843	0.0505
SIN5	0.07411	4.6868	0.0158	0.9874
COS67	-7.23206	4.4067	-1.6412	0.1045
SIN67	-9.30491	4.4084	-2.1107	0.0378
COS77	-1.81991	4.3369	-0.4196	0.6758
SIN77	-6.91688	4.3410	-1.5934	0.1149
COS123	-7.27176	3.9828	-1.8258	0.0715
SIN123	5.92535	3.9767	1.4900	0.1400
COS143	-11.25509	3.8120	-2.9526	0.0041
SIN143	-4.88520	3.8107	-1.2820	0.2034
IWS[/D(1)]	-0.48453	0.0583	-8.3145	< .0001
IWS[/D(1)] Den1	0.50785	0.0841	6.0388	< .0001
Model Variance (sigma squared)	3457	.	.	.

Table 4.2: Parameter Estimates for champion model

Finally, we have the parameter estimate as in Table 4.2. Although some trigonometric variables such as COS77 are not quite significant, the model has a desirable performance and successfully captured the majority of extreme values and most variations among different periods. Therefore, in conclusion, with Cyclical Trend with ARMA(1,1) and TF Model, we finally achieve a preferable fit towards our air pollution series.