Chen Wang

Dr. Richard H. Hartwell

MSSP6070 Practical Programming for Data Science

22 December 2022

Report for Beijing PM2.5 Variation between 2013 and 2017

Executive Summary

China is rated as the 'high risk' area for PM2.5 concentration according to Lim et al.'s' work, which fine particular matter is a Group 1 carcinogen in 2013 as reported by World Health Organization. Facing the threat, China triggered the toughest-ever clean air policy in 2013 (Zhang, Zheng, et al.) and broad attention shifted to this topic.

This report employs Beijing's hourly PM2.5 data from 11 monitoring sites along with other 11 types of meteorological data from 2013 Mar. to 2017 Feb. It attempts to explore the spatial and temporal variations of PM2.5 pollutant, especially the relationship with meteorological data and seasonal trends. The analyses bring out valuable insights for controlling pollution. First, the eastern wind direction exacerbates PM2.5 pollution in Beijing and the increased wind speed facilitates the spread of pollution and improves the level of air quality. Second, the situation worsens in the center of Beijing city rather than the suburbs. Third, other pollutants, such as SO2 and NO2 could indirectly contribute to the formation of PM2.5 through complex chemical reactions. Finally, the volatility level of PM2.5 has clear monthly and daily patterns which are captured in the SARIMAX time series model.

I. Introduction

With rapid urbanization and great energy consumption, developing countries especially China and India, have been exposed to a high concentration of PM2.5 since 1998 (Lim et al.). According to US EPA, PM2.5 describes an atmospheric particulate matter with a diameter lower than 2.5 micrometers. PM2.5 pollution not only reduces outdoor visibility but deteriorates human respiratory systems (Lin et al.), leading to a reduced average life expectancy of 1.2-1.9 years in polluted countries in Asia and Africa (Lim et al.). Various studies found that severe pollution events are influenced by a mixture of emissions from traffic, industries, and constructions and meteorological factors (Wang et al.; Zhang et al.). Wang and colleagues reported that the consideration of meteorology and atmospheric chemistry promotes the model performance that accounts for 30% of monthly PM2.5 variation on average.

This case report aims to explore the patterns of PM2.5 variability and build two models for predicting pollution events incorporating weather factors, pollutants, and non-stationary characteristics. The analysis is based on the hourly datasets of air pollutants data from March 2013 to Feb 2017 from 12 nationally-controlled monitoring sites (see Fig. 1). It further matches with nearby weather stations from the China Meteorological Administration. Notably, this research starts with data from March 2013. This is because the operation of the monitoring networks in Beijing started in January 2013 and there were many missing values between January to February 2013. Besides, the winter season is not separated into two years, allowing us to draw conclusions about the seasonal changes.

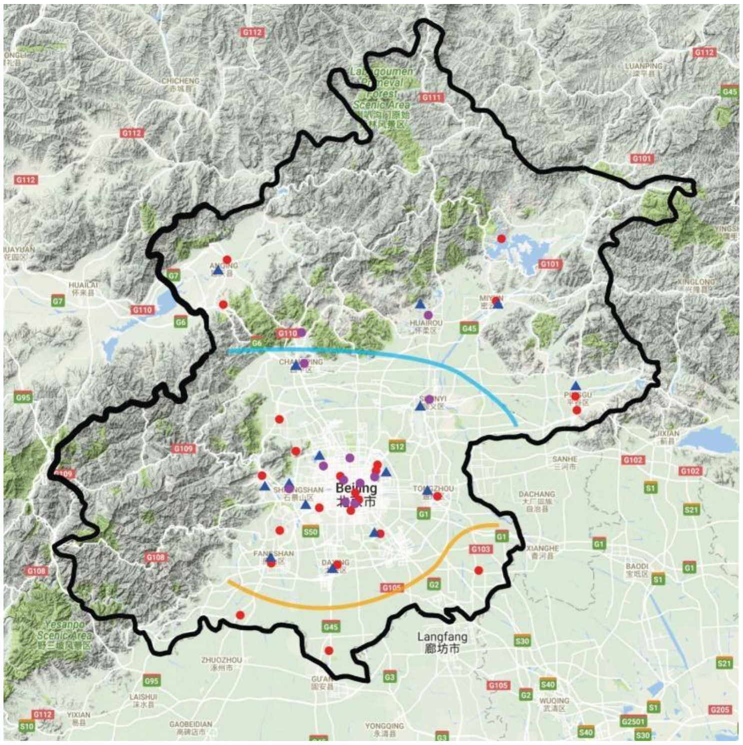


Figure 1. Locations of the 36 air-quality monitoring sites and 15 meteorological sites in Beijing. The purple dots mark the 12 monitoring sites used in this report.

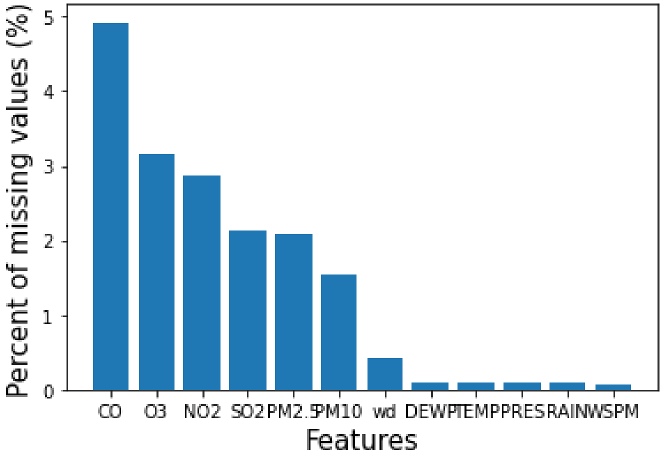
II. Methods

Data

The time series dataset consists of 420768 observations and 19 columns in total. The air pollutants dataset contains our response variable PM2.5, and other pollutants (**PM10**, **SO2**, **NO2**, **CO**, and **O3**). The meteorological data include 6 variables: temperature (**TEMP**), pressure (**PRES**), dew point temperature (**DEWP**), precipitation (**RAIN)**, wind direction (**wd**), and wind speed (**WSPM**). The dew point is the temperature the air needs to be cooled to in order to achieve a humidity of 100%, expressing humidity (Wood). Other variables include the station name **(station)** that monitors the air quality. We are interested in 1) whether the observed level of PM2.5 is associated with other pollutants emissions and meteorological data; 2) the effect of seasonal trends.

First, we deal with 74027 missing observations in the dataset. According to Fig 2., the missing values only constitute a small proportion, with CO being the variable with the largest missing value ratio (4.9%). For regression analysis, we impute the missing values all by their median values to avoid the influence of extreme values.

Figure 2. Percent missing data by feature



For the response variable PM2.5, it is highly right-skewed (skewness = 2.05) as plotted in the histogram of Fig 3. and boxplot of Fig 4. Thus, a log transformation for PM2.5 is appropriate for later multivariable linear modeling. After the transformation, the skewness is -0.41 and the distribution approximates normal.

Figure 3. PM2.5 Hist Plot

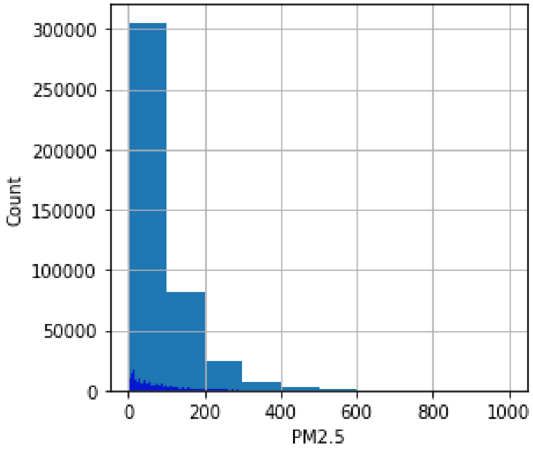


Figure 4. PM2.5 Box Plot

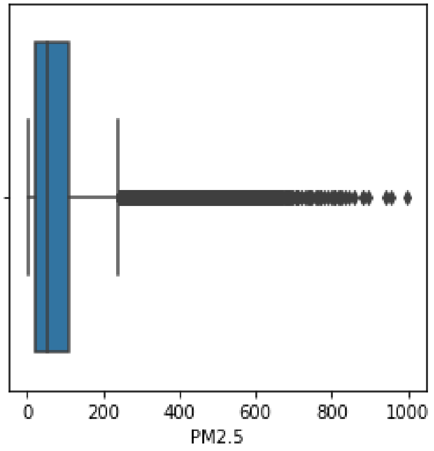


Figure 5. logPM2.5 Hist Plot

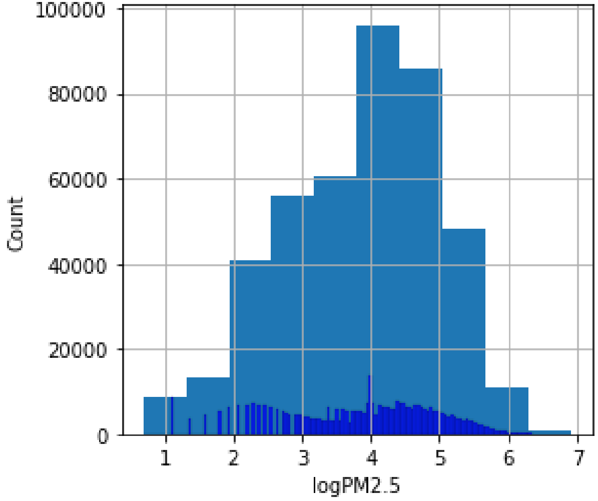
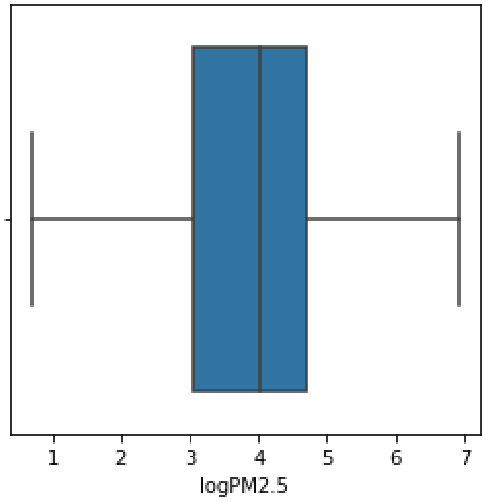


Figure 6. logPM2.5 Box Plot



Analysis

First of all, a boxplot is plotted for categorical variables, space and wind direction intuitively. As for quantitative variables, a multivariable linear regression is adopted to examine the effect of factors and their fitness level by R^2 and Root Mean Square Error. The Ordinary least square form is specified as follows:

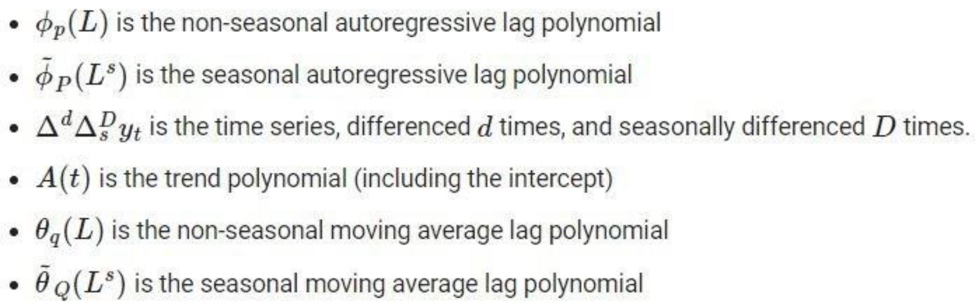
$$Y_{it} = \beta_0 + \beta_1 x_{1i} \cdots + \beta_k x_{ki} + \varepsilon_i$$

where $Y_{it\ i}$s the vaccination per capita for each state i at time t, β_0 is the intercept coefficient, β_1 to β_t represent the coefficient before other regression variables x, and ε is the error term.

Finally, we use a time series model of SARIMAX, stands for Seasonal Autoregressive Integrated Moving Average Exogenous model) (Chatfield).

$$\phi_p(L)\tilde{\phi}_P(L^s)\Delta^d\Delta_s^D y_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\epsilon_t$$

Above is a formula for the SARIMAX model where:



- $\phi_p(L)$ is the non-seasonal autoregressive lag polynomial
- $\tilde{\phi}_P(L^s)$ is the seasonal autoregressive lag polynomial
- $\Delta^d\Delta_s^D y_t$ is the time series, differenced d times, and seasonally differenced D times.
- $A(t)$ is the trend polynomial (including the intercept)
- $\theta_q(L)$ is the non-seasonal moving average lag polynomial
- $\tilde{\theta}_Q(L^s)$ is the seasonal moving average lag polynomial

III. Results

Descriptive statistics

Table 1. summarizes the descriptive statistics for the whole sample, there are n = 420768 observations in total. For variables TEMP, PRES, DEWP, RAIN, and WSPM, the mean is close to the median value, suggesting no extreme values pulling up the mean. Their values of standard deviation are also small at around 10. In contrast, PM10 has higher mean values than the median with clear right-skewed patterns. This is suggested by its high standard deviations (91). Furthermore, the observations are averagely distributed by 12 stations with 35064 observations each station.

Table 1. Descriptive Statistics

Variable	Unit	Min	Max	Mean	Median	S.D.
PM2.5	ug/m^3	2	999	79.27	55	80.05
PM10	ug/m^3	2	999	104.25	82	91.1
SO2	ug/m^3	0.28	500	15.64	7	21.45
NO2	ug/m^3	1.02	290	50.41	43	34.64
CO	ug/m^3	100	10000	1214.5	900	1134
O3	ug/m^3	0.21	1071	56.98	45	55.8
TEMP	degree Celsius	-19.9	41.6	13.53	14.5	11.43
PRES	hPa	982	1043	1010.7	1010	10.46
DEWP	degree Celsius	-43.4	29.1	2.49	3	13.78
RAIN	mm	0	72.5	0.06	0	0.82
WSPM	m/s	0	13.2	1.72	2.2	1.24

Then, we consider the correlation between multiple variables, as shown in Fig 7.

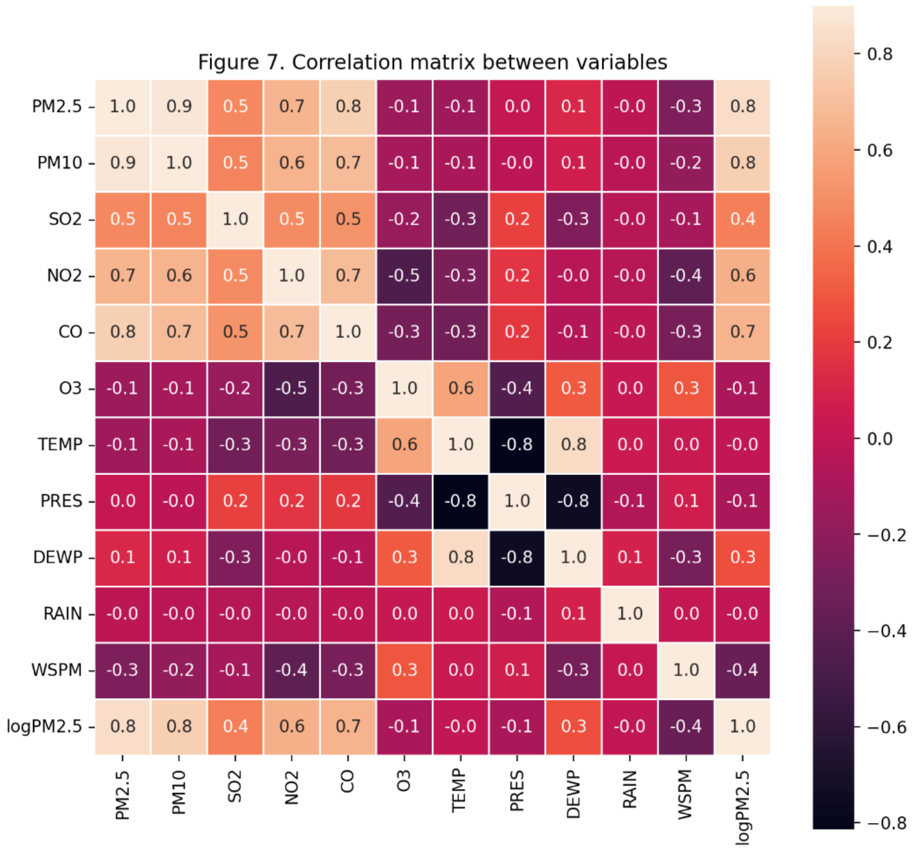


Figure 7. Correlation matrix between variables

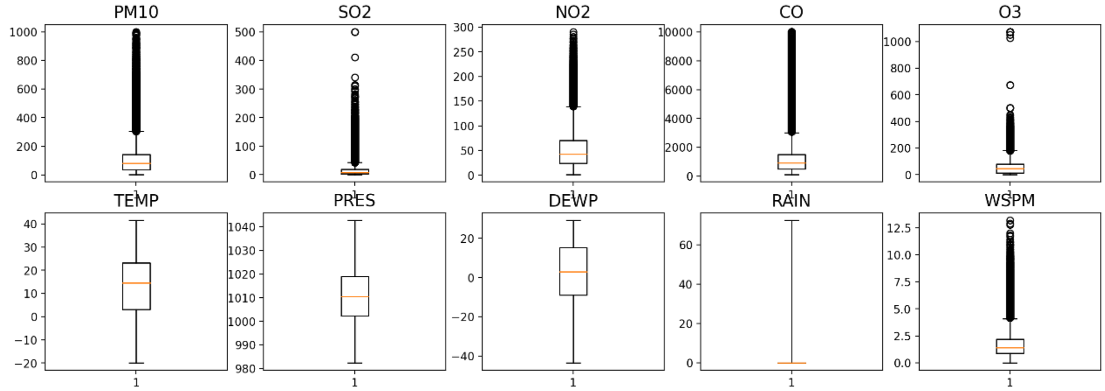
We only consider the main variable in the log scale. In this preliminary analysis, PM10 (0.8), NO2 (0.6), and CO (0.7) are highly correlated (>0.5) with our response variable. Moreover, several variables may encounter multicollinearity issues, with coefficients larger than 0.5. To eliminate multicollinearity concern, this report adopts VIF and found that TEMP (16.27), PRES (42.32), and PM10 (11.54) have a VIF value above 10. Thus, these variables will not be included in the regression analysis.

Outliers Examination

We already identify the distribution of the response variable. In this section, other dependent variables are examined through the boxplot.

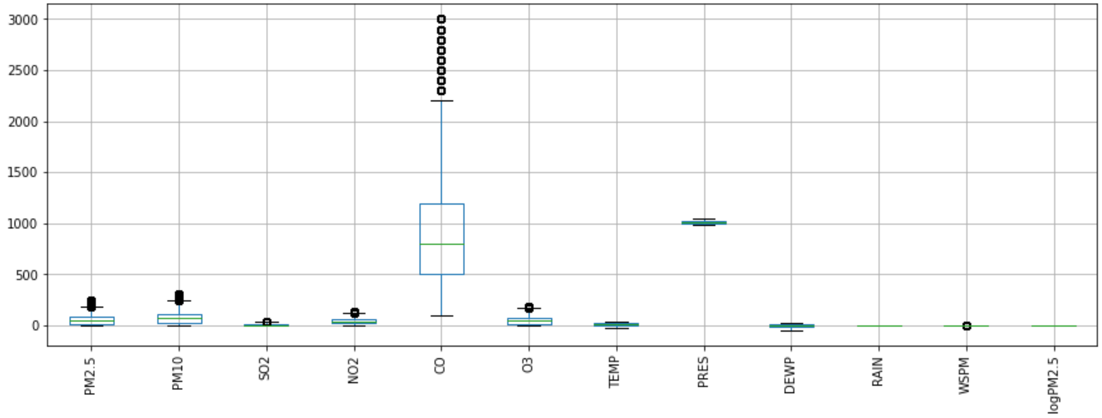
From Figure 8., PM10, SO2, NO2, CO, O3, and WSPM have extremely high values which exceed the upper limit (75%) and pull the mean values above the median.

Figure 8. Boxplot of other dependent variables



For the outlier treatment, IQR method is used to measure the variability by dividing the dataset into quantiles and calculating the range between the first and third quantiles. Those who fall below IQR -1.5 and above 1.5 are outliers. In this case, our dataset identifies more outliers. Notably, it is not appropriate to delete the values identified by IQR methods in the preliminary model. Instead, this report would compare the regression before and after the outliers' treatment to capture the improvement in model fitness. 27674 observations are identified by IQR method.

Figure 9. Boxplot of all numeric variables after outlier treatment



Categorical Variable Analysis

Figure 10. Boxplot of the PM2.5 by wind direction

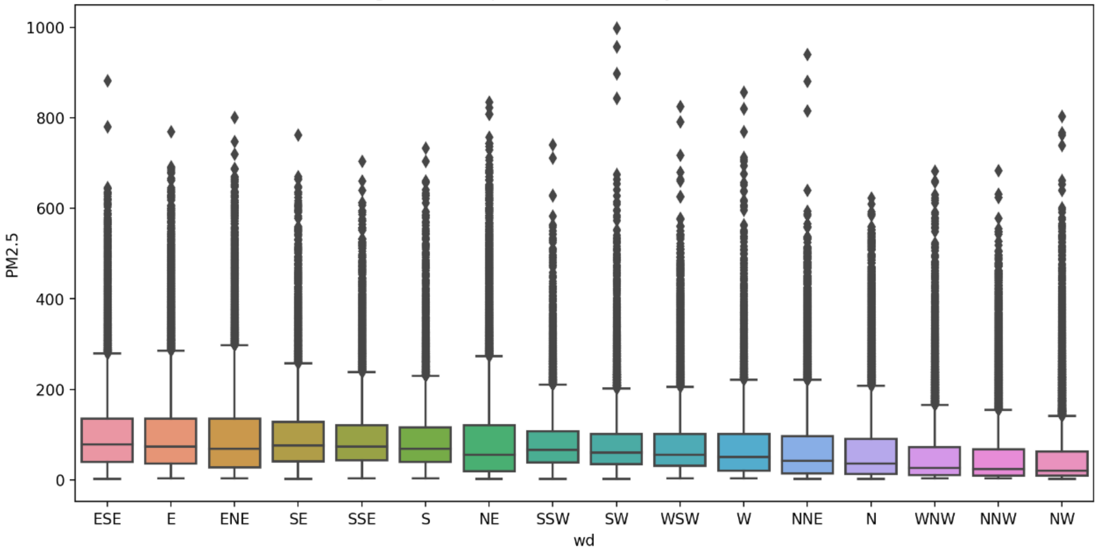
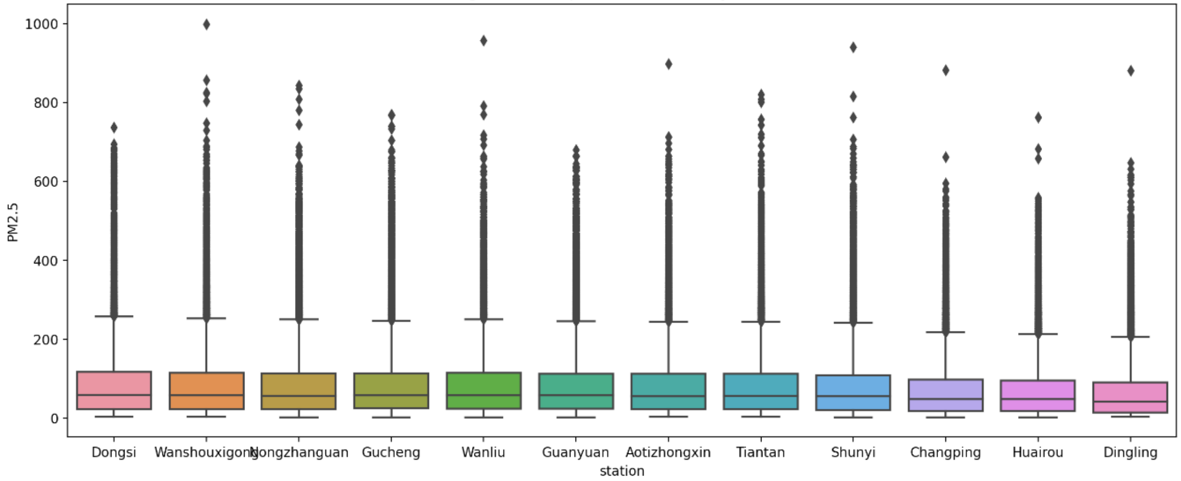


Figure 11. Boxplot of the PM2.5 by station

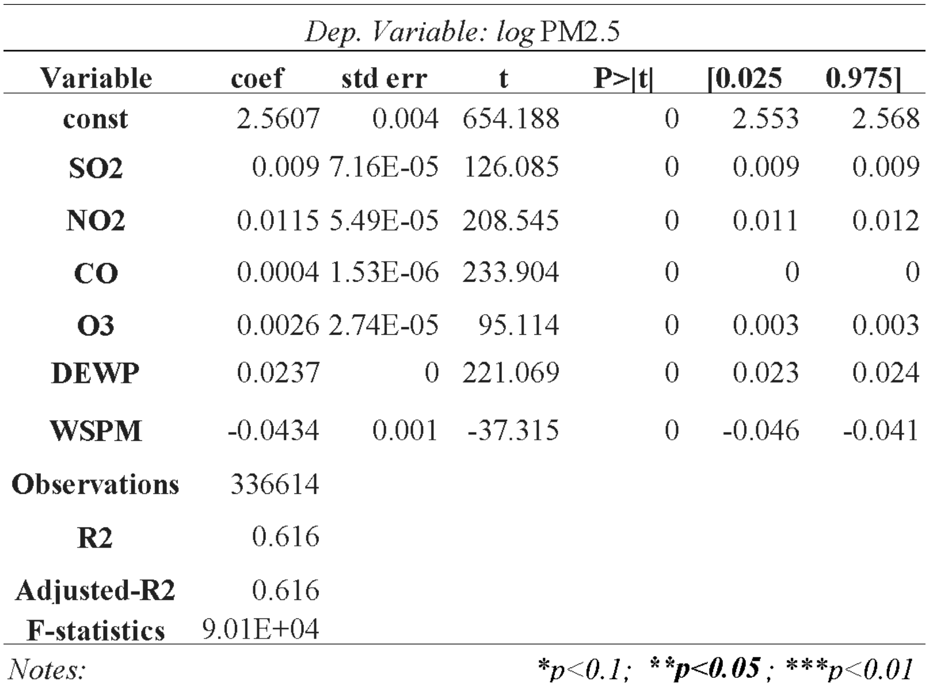


According to Fig 10, the east wind direction (both eastern-south and eastern-north) increases the occurrence of PM2.5 pollution. This owes to the fact that the northwest side of Beijing is surrounded by mountains that hinder the spread of contaminants. Meanwhile, the Hebei Province with heavy industrial bases is to the southeast of Beijing. In Fig 11, aside from the wind direction's influence, monitoring states in the urban area are more likely to be affected by PM2.5 than those in the suburbs. This is consistent with theoretical explanations that pollution is driven by commercial, residential, traffic, and construction activities in the urban areas.

Base Regression Model

First, a multivariable model is used to examine the effect of a mixture of pollutants and weather influences on the level of PM2.5.

Table 2. Regression results for linear OLS model



Dep. Variable: log PM2.5

Variable	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5607	0.004	654.188	0	2.553	2.568
SO2	0.009	7.16E-05	126.085	0	0.009	0.009
NO2	0.0115	5.49E-05	208.545	0	0.011	0.012
CO	0.0004	1.53E-06	233.904	0	0	0
O3	0.0026	2.74E-05	95.114	0	0.003	0.003
DEWP	0.0237	0	221.069	0	0.023	0.024
WSPM	-0.0434	0.001	-37.315	0	-0.046	-0.041
Observations	336614					
R2	0.616					
Adjusted-R2	0.616					
F-statistics	9.01E+04					

Notes: *$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$*

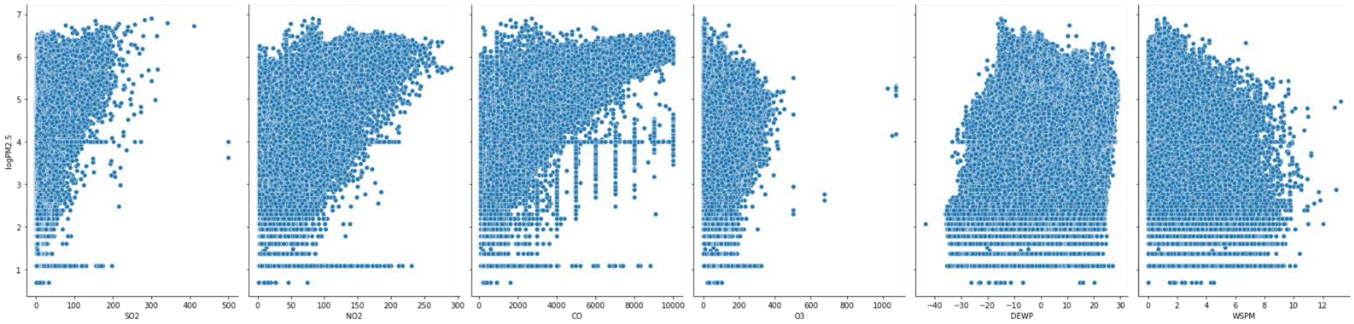
From Table 2, the model fitness is captured by the R-squared that only 61% of the variance in PM2.5 is explained by the inclusion of six independent variables. This does not indicate strong fitness and will be discussed later. Considering the pollutants, all three pollutants are positively correlated with the concentration of PM2.5 at a 5% level significance, though the magnitude is small. The coefficient before NO2 is 0.0114, showing that with a 1 ug/m^3 increase in NO2 level, the PM2.5 concentration has an associated 1.15 ug/m^3 increase of PM2.5 (calculated by (exp(0.0114)-1)*100). Concerning the meteorological factors, the relationship between the dew point temperature and PM2.5 is significantly positive at a coefficient of 0.0237. However, wind speed seems to decrease the concentration of PM2.5 with a negative coefficient (-0.04).

Initially, these interpretations are consistent with previous assumptions. To evaluate the model, the dataset has been divided into both train (80%) and test (20%) datasets. The RMSE is 0.697.

Linearity

From Fig13., PM2.5 are none of the independent variables that form a linear shape with great noises, although SO2, NO2, CO show clearer patterns than other variables. Thus, linear regression might not be the best for fitting the model. To eliminate the issue, advanced techniques such as machine learning are more appropriate.

Figure 13. Scatterplot of all variables included (SO2, NO2, CO, O3, DEWP, WSPM)



Homoscedasticity

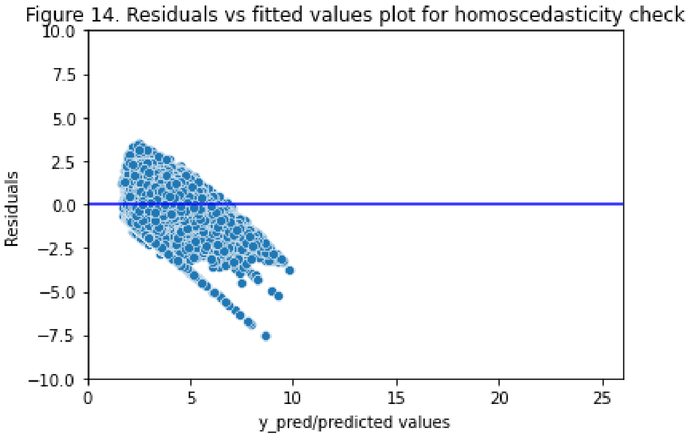


Figure 14. Residuals vs fitted values plot for homoscedasticity check

By checking Fig 14., it is apparent that the error terms against the predicted values fluctuate a lot and are not centered around 0. Thus, the assumption of homoskedasticity is violated as well.

Normality of residuals

According to the histogram and QQ plot from Fig.13 and Fig. 14, the distribution of residuals seems to be normal with skewness of -0.565. However, the p-value from the Kolmogorov-Smirnov test (Justel et al.), a method valid for large observations reaches 0, suggesting that the residuals are not normally distributed.

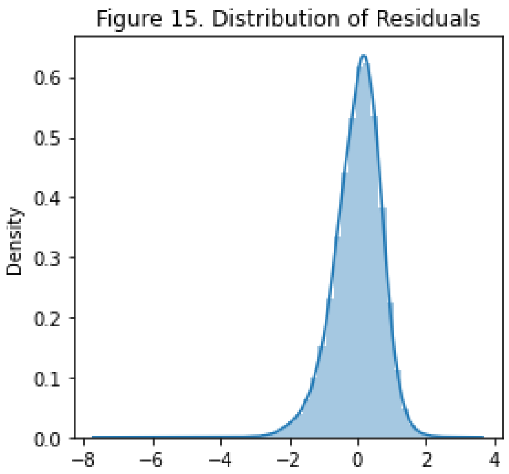


Figure 15. Distribution of Residuals

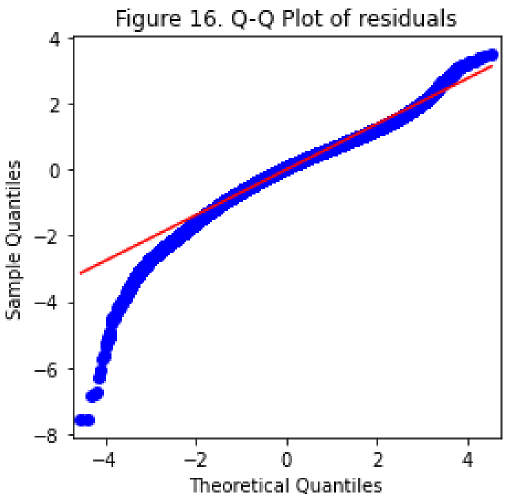
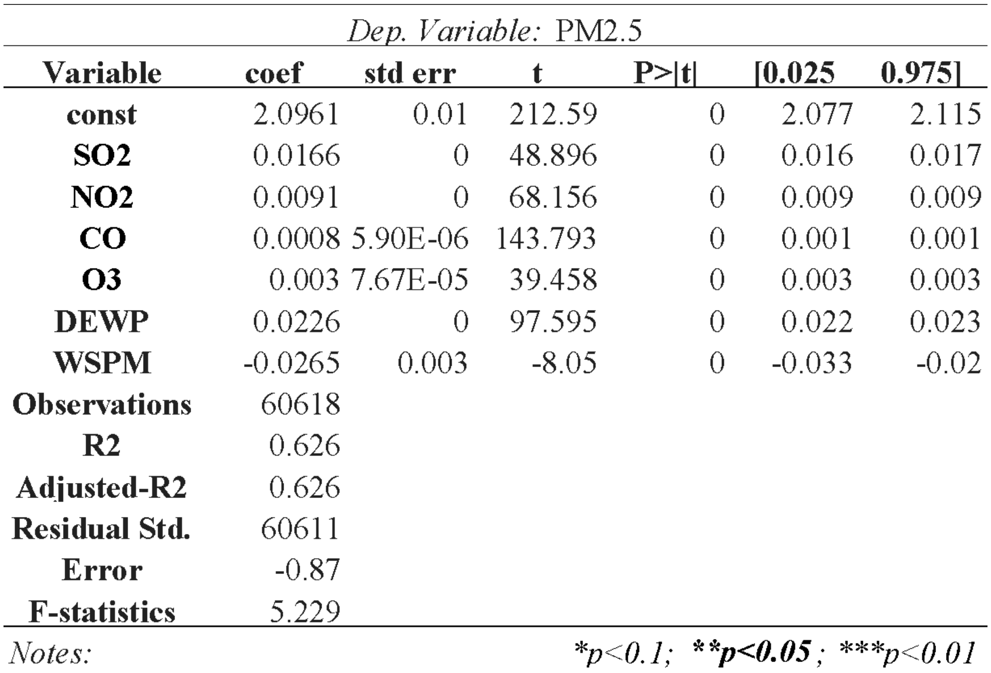


Figure 16. Q-Q Plot of residuals

Outliers Treatment

After trimming the outliers identified by the IQR method, we have 393094 observations in total. The significance of R^2 does not improve too much, increasing slightly from 0.616 to 0.628, and the magnitude of the effect of SO2, CO and O3 improved mildly compared to Table 2. The RMSE after the outlier treatment decreases to 0.627, which implies the quality of the model improves. According to diagnosis tests, the model still not meet the requirements of three classic assumptions.

Table 3. Regression results for linear OLS model After outlier treatment



Dep. Variable: PM2.5

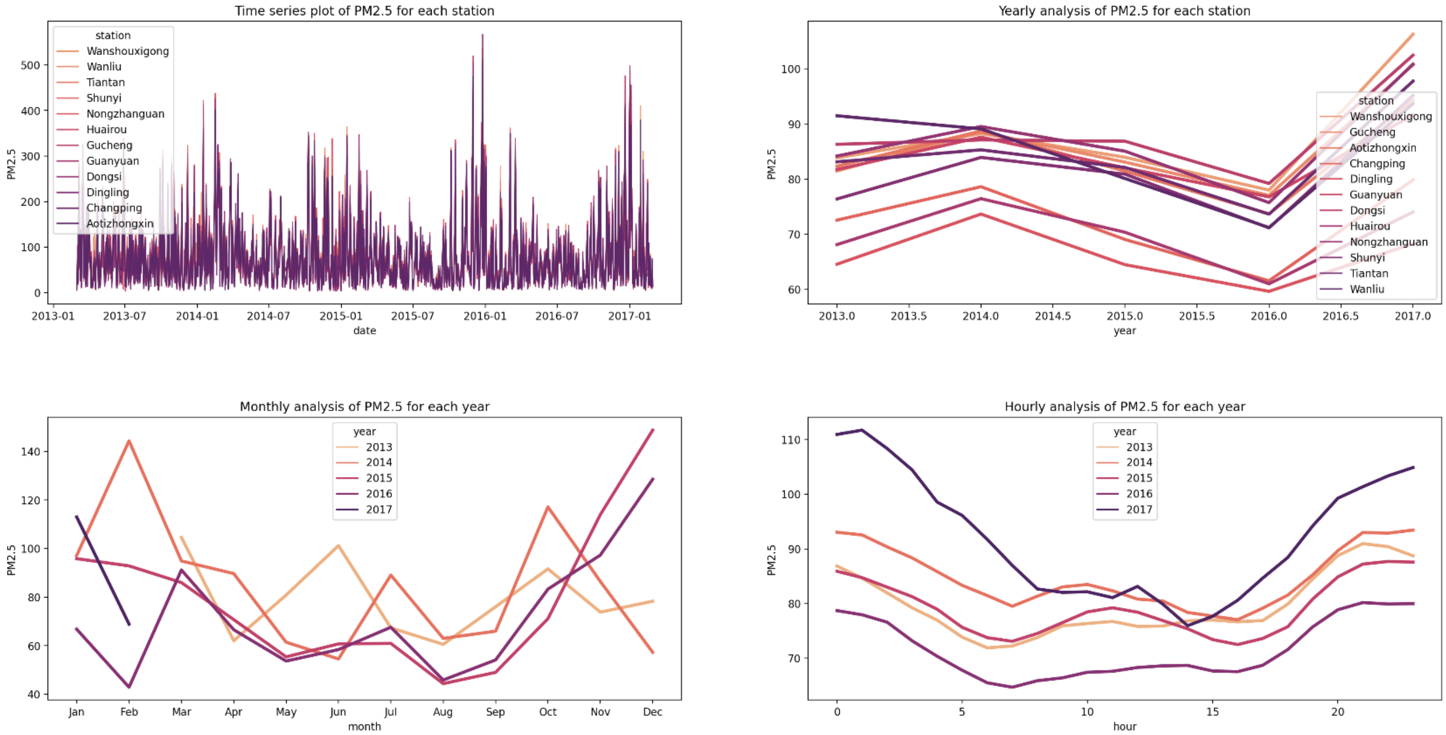
Variable	coef	std err	t	P>|t|	[0.025	0.975]
const	2.0961	0.01	212.59	0	2.077	2.115
SO2	0.0166	0	48.896	0	0.016	0.017
NO2	0.0091	0	68.156	0	0.009	0.009
CO	0.0008	5.90E-06	143.793	0	0.001	0.001
O3	0.003	7.67E-05	39.458	0	0.003	0.003
DEWP	0.0226	0	97.595	0	0.022	0.023
WSPM	-0.0265	0.003	-8.05	0	-0.033	-0.02
Observations	60618					
R2	0.626					
Adjusted-R2	0.626					
Residual Std.	60611					
Error	-0.87					
F-statistics	5.229					

Notes: **p<0.1;* *****p<0.05***; ****p<0.01*

SARIMAX Model

Since three linear model assumptions are not justified, I adopt one advanced times series model for accurate prediction. Different from the ARIMA model, the SARIMAX model is more powerful to relate to the seasonal part of the time series dataset but also deals with its non-seasonal part.

Figure 12. Time series plot of PM2.5, yearly, monthly and hourly analysis



By decomposing the dataset on a yearly, monthly, and hourly basis, we can see from Fig. 12 that all the stations follow similar patterns:

- The overall PM2.5 pollution jumped shortly after a slump in the Year 2016;
- PM2.5 hits the lowest level in the summer period which has the highest temperature and humidity, but bounces back in the winter;
- The noon has the lowest level of PM2.5 concentration, in contrast to the night, which period becomes the highest level.

Using the Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, the dataset is stationary (d = 0) and does not require differencing. Next, to fit the model with different (p, d, q) * (P, D, Q), an iteration model is used to find a combination of (p, d, q) * (P, D, Q) with the smallest AIC, which estimates the quality of the model and obviate the overfitting problem.

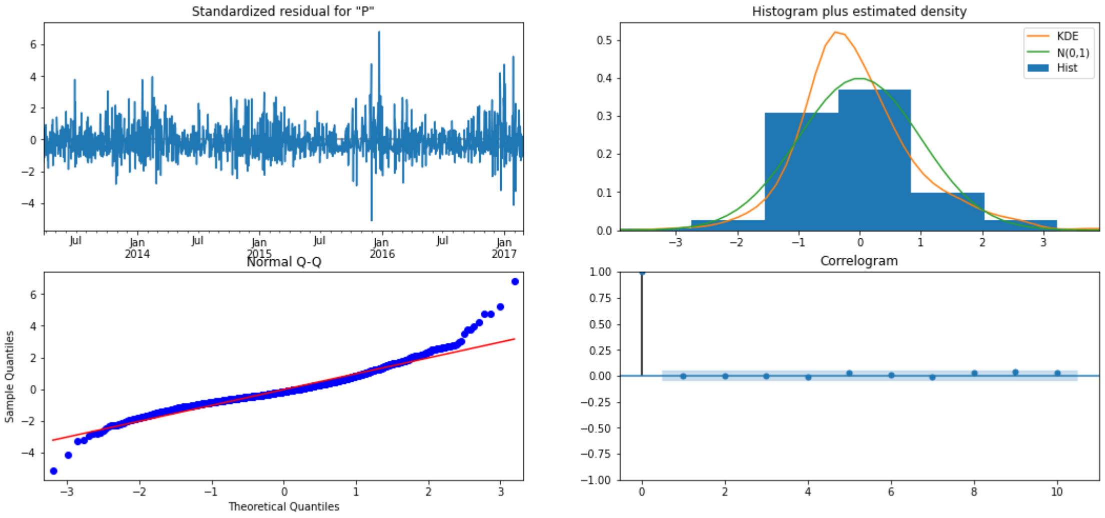
Table 4. SARIMAX Results

```
==============================================================================
Dep. Variable:                          PM2.5   No. Observations:         1461
Model:       SARIMAX(1, 0, 1)x(0, 1, 1, 12)   Log Likelihood       -7760.866
AIC:                                15529.732   BIC                  15550.807
Covariance Type:                          opg
==============================================================================
                   coef    std err        z      P>|z|     [0.025     0.975]
------------------------------------------------------------------------------
ar.L1              0.37      0.028   13.489      0.000      0.318      0.426
ma.L1              0.31      0.034    9.010      0.000      0.239      0.372
ma.S.L12          -1.00     29.385   -0.034      0.973    -58.593     56.593
sigma2          2819.54   8.29e+04    0.034      0.973    -1.6e+05   1.65e+05
==============================================================================
Ljung-Box (L1) (Q):                 0.00   Jarque-Bera (JB):         1109.54
Prob(Q):                            0.99   Prob(JB):                    0.00
Heteroskedasticity (H):             1.45   Skew:                        0.83
Prob(H) (two-sided):                0.00   Kurtosis:                    6.97
```

The SARIMAX(1, 0, 1)x(0, 1, 1, 12) with AIC value of 15529.7 is the best combination. From Table 4., the term of ma.S.L 12 is not significant at 5% level, but ar.L1 and ma.L1 have strong explanatory power. Examining four diagnosis plots in Fig 13.

- The mean of the residuals is around 0, and there are no obvious patterns in residuals. This is verified by Ljung-Box test (p-value = 0.99 > 0.5) that no correlations in residuals.
- The distribution of residuals is normal overall. Also from QQ-plot, most of data lie on the 45°line, suggesting the normality of residuals. However, the Jarque-Bera test (p-value =0) shows the residuals are not normally distributed.
- In the correlogram or ACF plot, 95% of correlations for lag greater than one should not be significant (inside the blue area).

Figure 13. Diagnosis summary of SARIMAX model



Finally, for model evaluation, the mean absolute error (MAE) is 48, a bit large. From Fig 14., the yellow lines are our prediction according to the SARIMAX model. This is within the 95% confidence interval of the actual variation of PM2.5 since 2016 Aug 31st.

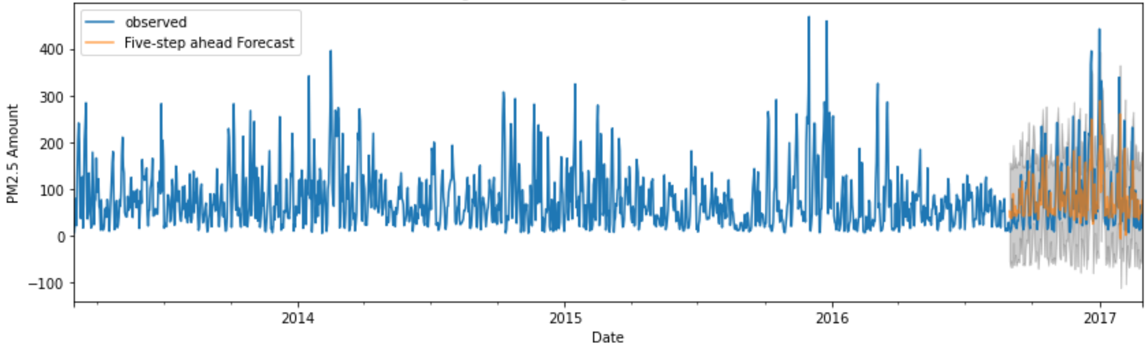


Figure 14. Forecasting 6 months ahead

IV. Discussion and Conclusion

One limitation is the methodology of multivariable regression analysis. The model fitness does not achieve our expectations and three classic assumptions fail even after trimming outliers. According to the diagnosis plots, it might be the case that the relationship between PM2.5 concentration and other independent variables is not linear, the pollution values have nonstationary patterns such as seasonality, or there are other characteristics that are not captured. Moreover, our SARIMAX model can be improved further by including the exogenous variables and modifying the parameters. Aside from these models, LSTM (Long Short Term Memory) technique is powerful to improve the capacity of prediction in the time series dataset, while we do not conduct such techniques here due to its complexity.

While this report has limitations, it provides valuable policy suggestions for vulnerable developing countries like China to improve their air quality. First of all, countries should establish multiple monitoring sites to collect real-time pollution data for prediction and diagnosis. Second, meteorological variables such as humidity and wind speed are important predictors of the dramatic variability of PM2.5 concentration, rather than the process of emission, though emission is significant for PM2.5 reduction in the long term.

Works Cited

Chatfield, Chris. *The Analysis of Time Series*. Chapman and Hall/CRC, 2003, https://doi.org/10.4324/9780203491683.

Justel, Ana, et al. “A Multivariate Kolmogorov-Smirnov Test of Goodness of Fit.” *Statistics & Probability Letters*, vol. 35, no. 3, Oct. 1997, pp. 251–59, https://doi.org/10.1016/s0167-7152(97)00020-5.

Lim, Chul-Hee, et al. “Understanding Global PM2.5 Concentrations and Their Drivers in Recent Decades (1998–2016).” *Environment International*, vol. 144, Nov. 2020, p. 106011, https://doi.org/10.1016/j.envint.2020.106011.

Lin, Yaolin, et al. “A Review of Recent Advances in Research on PM2.5 in China.” *International Journal of Environmental Research and Public Health*, vol. 15, no. 3, Mar. 2018, p. 438, https://doi.org/10.3390/ijerph15030438.

US EPA. “Particulate Matter (PM2.5) Trends.” *US EPA*, 19 July 2016, www.epa.gov/air-trends/particulate-matter-pm25-trends.

Wang, Lili, et al. “The Influence of Climate Factors, Meteorological Conditions, and Boundary-Layer Structure on Severe Haze Pollution in the Beijing-Tianjin-Hebei Region during January 2013.” *Advances in Meteorology*, vol. 2014, 2014, pp. 1–14, https://doi.org/10.1155/2014/685971.

Wang, ZiFa, et al. “Modeling Study of Regional Severe Hazes over Mid-Eastern China in January 2013 and Its Implications on Pollution Prevention and Control.” *Science China Earth Sciences*, vol. 57, no. 1, Dec. 2013, pp. 3–13, https://doi.org/10.1007/s11430-013-4793-0.

Wood, Lawrence A. “The Use of Dew-Point Temperature in Humidity Calculations.” *Journal of Research of the National Bureau of Standards, Section C: Engineering and Instrumentation*, vol. 74C, no. 3-4, July 1970, p. 117, https://doi.org/10.6028/jres.074c.014.

World Health Organization (WHO). “Evolution of WHO Air Quality Guidelines: Past, Present and Future.” *Apps.who.int*, World Health Organization Regional Office for Europe, 2017, apps.who.int/iris/handle/10665/341912.

Zhang, Qiang, et al. “Drivers of Improved PM2.5 Air Quality in China from 2013 to 2017.” *Proceedings of the National Academy of Sciences*, vol. 116, no. 49, Nov. 2019, p. 201907956, https://doi.org/10.1073/pnas.1907956116.

Zhang, Shuyi, et al. “Cautionary Tales on Air-Quality Improvement in Beijing.” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2205, Sept. 2017, p. 20170457, https://doi.org/10.1098/rspa.2017.0457.