Beijing PM2.5 Project, 2014 - 2020

Ying Cui
Renping Ge
Ziyu Huang
Yu Cao
Chaohui Li

Group 5

Contents

Data Preprocessing

Detect	Detect and Define the missing values
Transformation	Transform heavy skewed variables
Scaling	Rescaling real-valued numeric attributes
New Value	Change variable old value to new value
Create	Create daily, monthly data

Data Preprocessing - Missing value

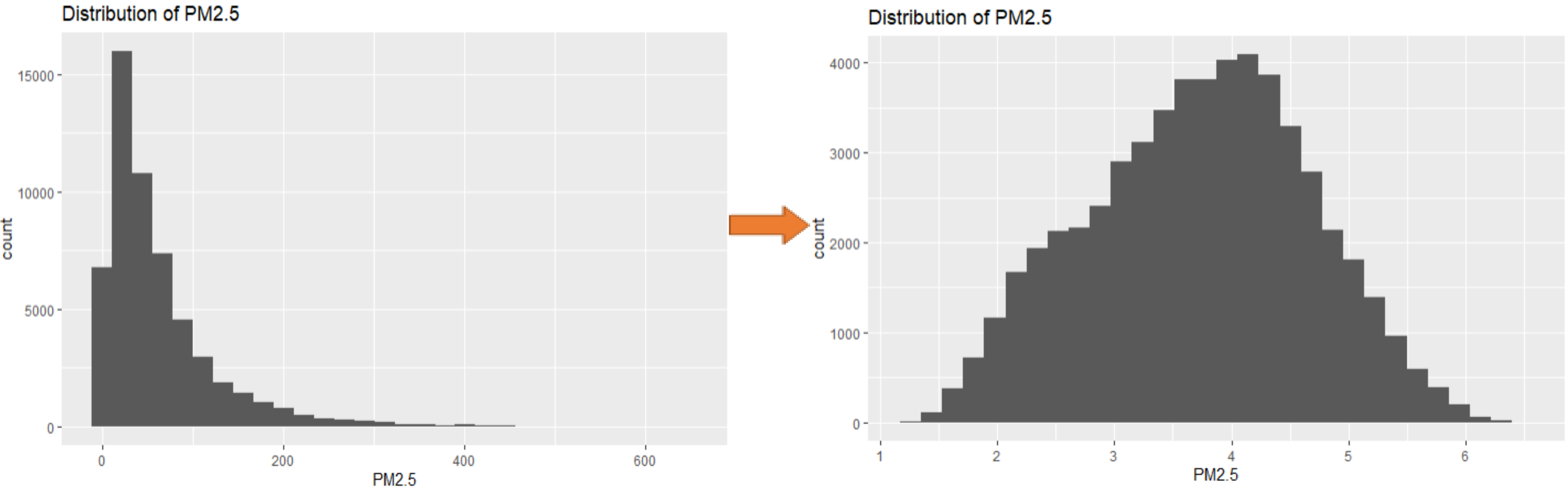
PM2.5	242
PM10	964
AQI	407
SO2	2390
NO2	2394
O3	2394
CO	3345
TEMP	4
DEWP	22
PRES	3917
Wd	8454
WSPM	2

Continuous Variables: PM10, AQI, SO2, NO2, O3, CO, TEMP, DEWP, PRES, WAPM

Categorical variable: Wd

Method: KNN Imputation

Data Preprocessing - Transformation log(1+X)



Data Preprocessing - Scaling

	WSPM	TEMP	DEWP	PRES	AQI	SO2	NO2	O3	CO
count	58444	58444	58444	58444	58444	58444	58444	58444	58444
mean	0.7655	0.2169	0.8029	0.9994	0.81243	0.6278	0.8694	0.75998	0.7515
median	0.8656	0.3418	0.6052	0.9999	0.64487	0.3290	0.7801	0.60888	0.5722
min	-0.8656	-2.5534	0	0.9754	0.09473	0.1122	0.0657	0.02149	0.1050
25%	0.2164	-0.6139	0.3026	0.9911	0.42141	0.1953	0.4960	0.25860	0.3618
75%	1.2984	1.1093	0.9079	1.0082	1.00403	0.7065	1.1542	1.03401	0.8767
max	2.2722	1.9534	6.0524	1.0m296	4.46636	15.1028	3.6152	4.21138	10.2398

Data Preprocessing - New variable value

Variable: Wind Direction

Old Value	No Wind	E	W	N	S	NE,ENE,NNE	SE,ESE,SSE	NW,NNW,WNW	SW,SSW,WSW
New Value	0	1	2	3	4	5	6	7	8

Data Preprocessing - Create daily, monthly data

Daily data

date	month	TEMP	DEWP	WSPM	PRES	PM2.5	PM10	AQI	SO2	NO2	O3	CO	wind
4/4/2014	4	14.59583	-8.83333	3.125	1015.958	3.440654	130.8253	76.59306	10.13898	39.48743	60.25103	1.04515	5
4/5/2014	4	12.80833	-9.11667	2.833333	1019.021	3.104168	85.06667	77.74171	7.742386	31.17574	74.20382	1.002381	8
4/6/2014	4	19.6	-9.21111	4.055556	1016.767	3.938102	124.0132	77.74419	12.55657	56.95446	51.92638	1.056676	8
4/8/2014	4	17.48571	5.328571	2.428571	1011.007	5.199433	233.3428	211.6756	48.87391	65.60904	141.861	1.843635	8
4/9/2014	4	20.9625	-1.0125	4	1012.896	4.725701	231.2291	202.0815	21.50838	49.57364	80.51685	1.4307	6
4/10/2014	4	15.43913	-2.73913	2.043478	1019.67	4.4837	388.8532	217.9295	26.90292	52.2497	47.72455	1.645786	6

Monthly data

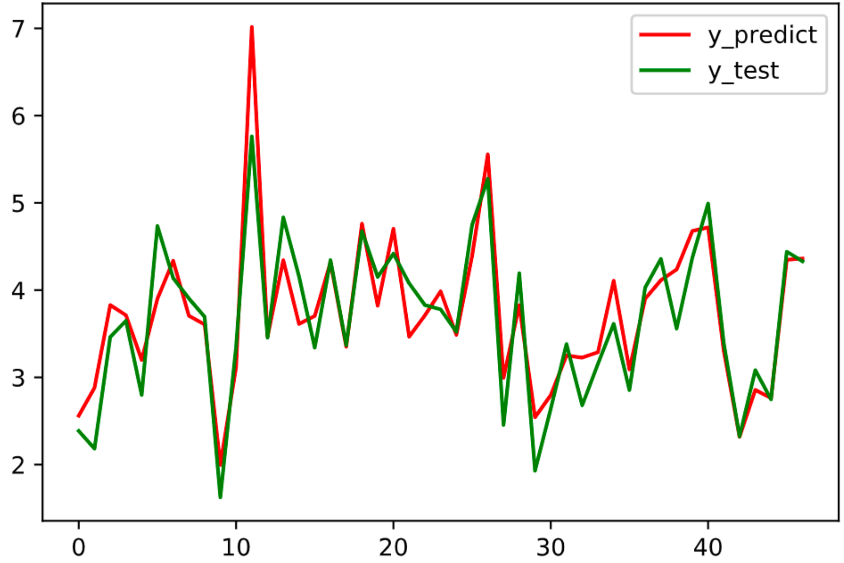
Year	month	TEMP	DEWP	WSPM	PRES	PM2.5	PM10	AQI	SO2	NO2	O3	CO	wind
2014	4	16.97423	2.30922	2.65721	1014.251	4.418991	152.6504	137.3981	19.86657	58.47947	71.51219	1.416803	8
2014	5	21.87677	6.562756	3.570652	1008.486	3.964887	116.4343	96.54859	15.68079	48.00267	87.87645	0.84034	8
2014	6	24.98291	15.74258	2.365546	1005.857	3.847853	77.86121	84.88091	8.739304	42.9775	93.89398	0.793652	8
2014	7	28.11161	19.22932	2.478754	1004.926	4.200687	109.3606	119.8755	7.994428	40.14899	96.10761	0.944006	8
2014	8	25.88286	17.75614	2.106613	1007.632	3.945834	97.13022	93.13668	6.805161	43.9799	89.711	0.865358	8

Regression Model

- Full Linear Regression Model
- Multicollinearity
- Lasso Regression

Full Linear Regression for Daily Data

	coef	Std err	P>|t|
TEMP	-47.2933	2.730	0.000
DEWP	43.5307	2.167	0.000
WSPM	43.5274	10.405	0.000
PRES	2.4535	0.057	0.000
SO2	-1.6932	1.405	0.228
NO2	28.9983	0.845	0.000
O3	7.9708	0.411	0.000
CO	251.8112	22.898	0.000
month	-33.8947	3.062	0.000
wind	-2.7379	4.434	0.537

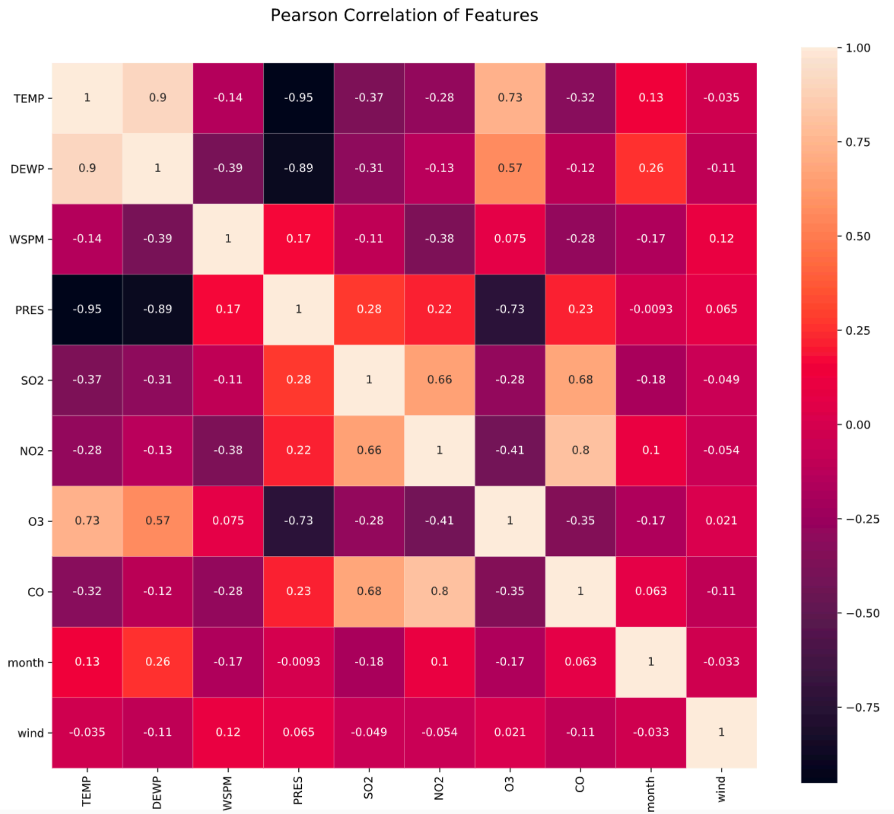


MSE:0.15108276403976398, R-square:0.8112430954158183

Cond. No. = 2.69e+03

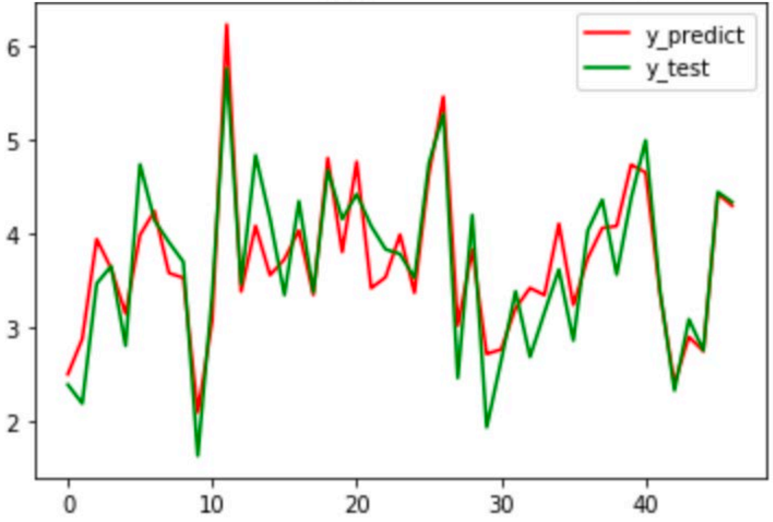
Multicollinearity Test

	VIF	features
0	32.4	TEMP
1	12.2	DEWP
2	11.9	WSPM
3	43.1	PRES
4	4.5	SO2
5	21.3	NO2
6	11.1	O3
7	10.4	CO
8	6.7	month
9	11.4	wind



Lasso Regression

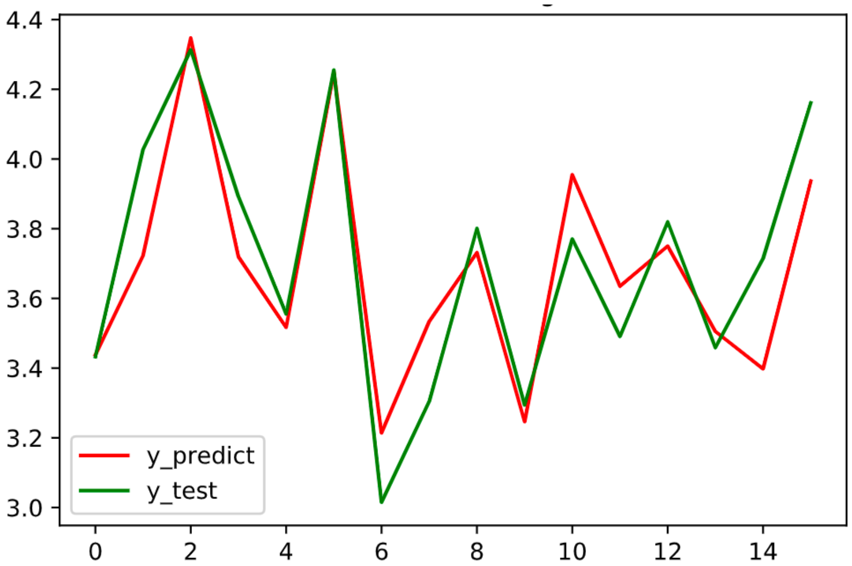
	coef	P>|t|	VIF
TEMP	-43.0866	0.000	14.0
DEWP	37.7111	0.000	4.5
SO2	4.188738	0.000	3.8
NO2	33.44839	0.000	8.2
O3	0.02911	0.000	6.9
month	-18.688	0.000	4.2



MSE:0.14810501066494958, R-square:0.8149633841808367

Full Linear Regression for Monthly Data

	coef	Std err	P>|t|
TEMP	-25.888	17.243	0.139
DEWP	24.9087	13.222	0.065
WSPM	-11.0534	89.632	0.902
PRES	2.6787	0.311	0.000
SO2	-2.0818	6.511	0.750
NO2	31.6646	3.706	0.000
O3	7.0600	2.278	0.003
CO	159.7193	123.869	0.203
month	-26.7558	8.994	0.004
wind	-48.3888	29.446	0.113

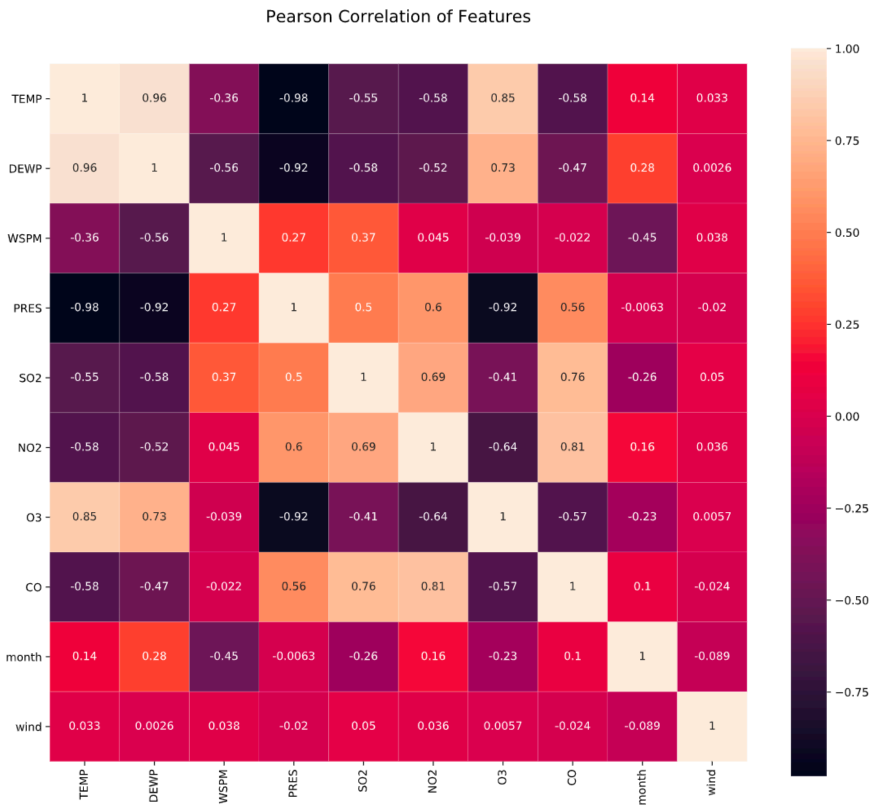


MSE:0.027279672017913547, R-square:0.7875313147539452

Cond. No. = 6.19e+03

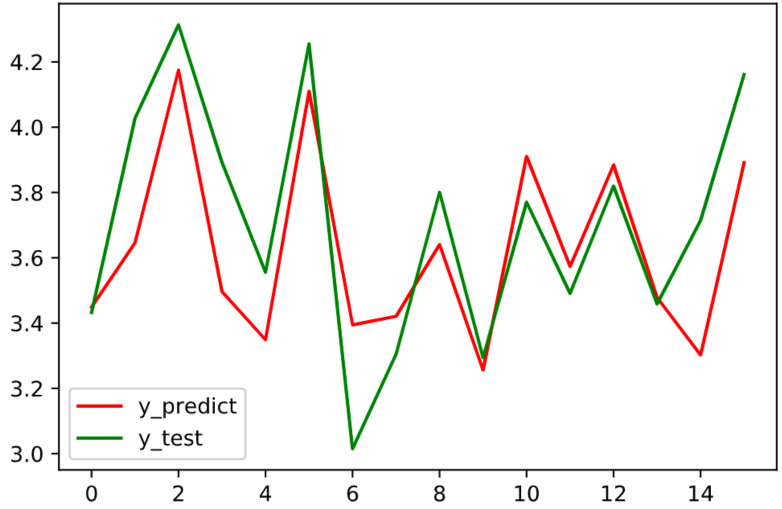
Multicollinearity Test

	VIF	features
0	239.3	TEMP
1	74.7	DEWP
2	132.8	WSPM
3	218.3	PRES
4	13.0	SO2
5	68.1	NO2
6	55.4	O3
7	46.4	CO
8	11.4	month
9	84.1	wind



Lasso Regression

	coef	P>|t|	VIF
DEWP	5.137299	0.000	2.5
SO2	4.149011	0.000	5.9
NO2	31.40919	0.000	22.6
O3	4.30546	0.000	4.9
month	-5.53937	0.024	7.6



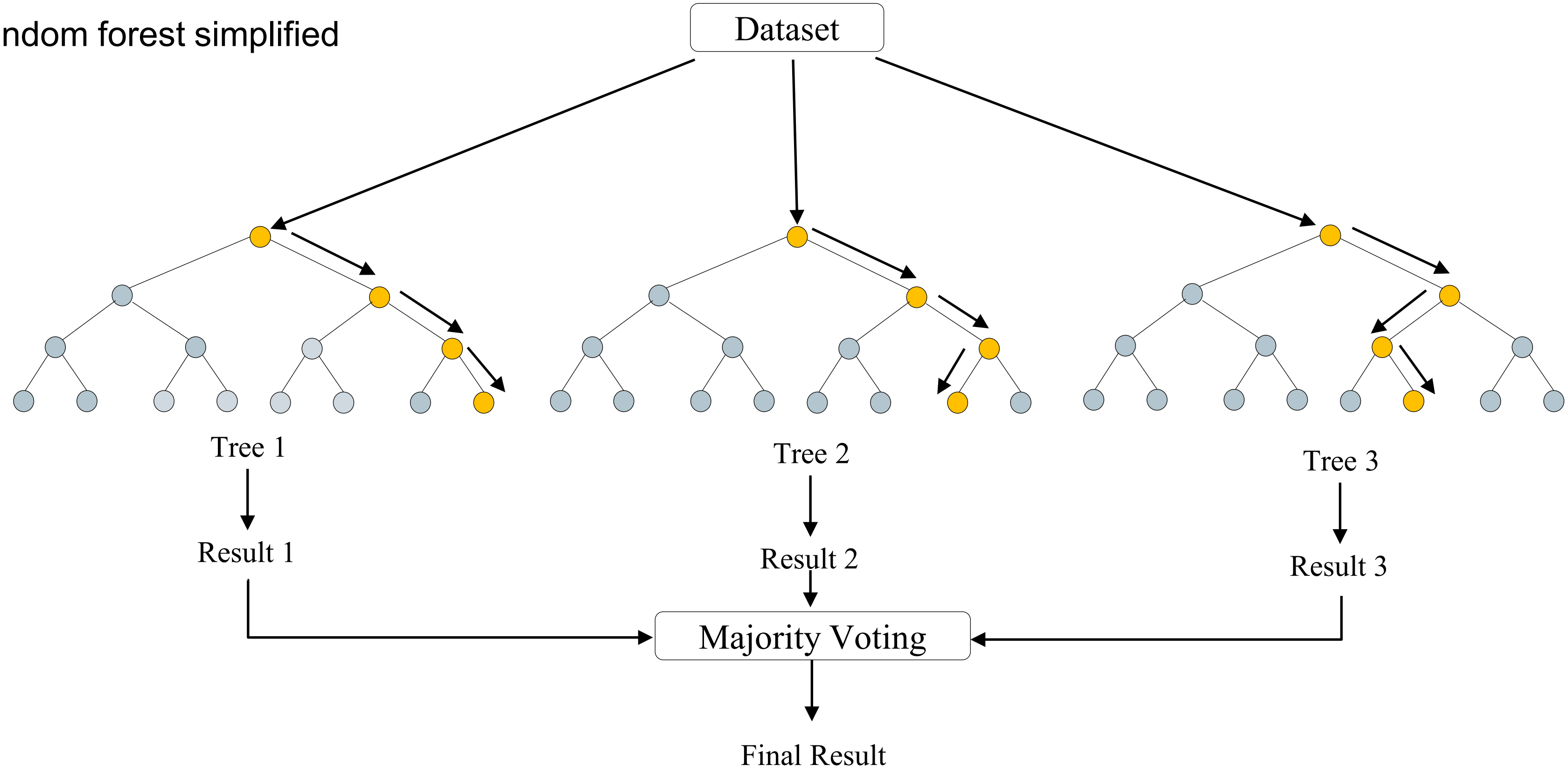
MSE:0.052626351175374786, R-square:0.5901178124067399

Random Forest

- Ensemble learning method
- Constructed of multitude decision trees

Random Forest

Random forest simplified

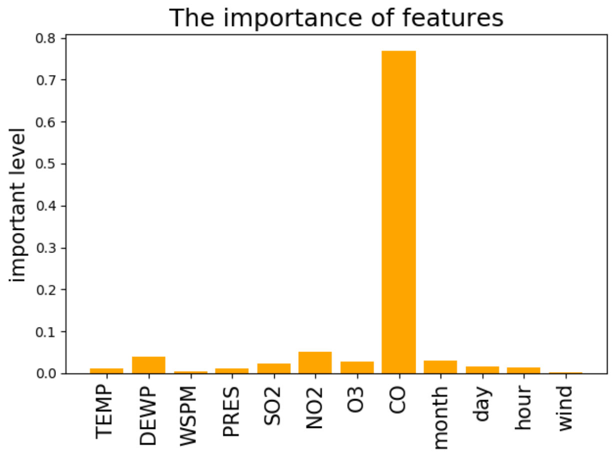


Random Forest

Hourly data

Pollution	Importance
CO	0.769
NO2	0.052
O3	0.028
SO2	0.023

Weather	Importance
DEMP	0.039
month	0.029
day	0.016
hour	0.013
PRES	0.012
TEMP	0.012
WSPM	0.005
wind	0.003

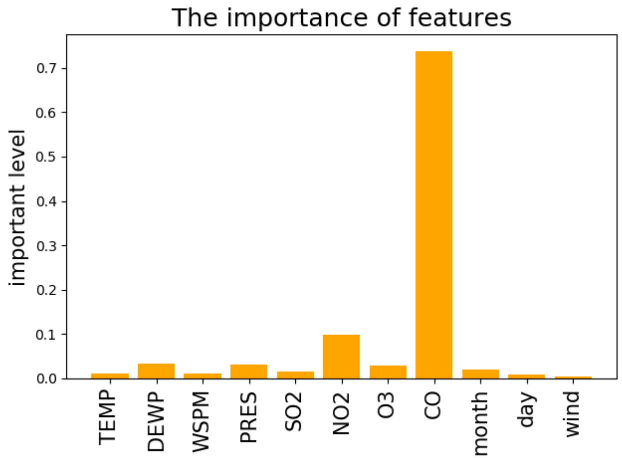


Random Forest

Daily data

Pollution	Importance
CO	0.738
NO2	0.098
O3	0.029
SO2	0.015

Weather	Importance
DEMP	0.034
PRES	0.031
TEMP	0.012
WSPM	0.01
month	0.02
day	0.009
wind	0.004

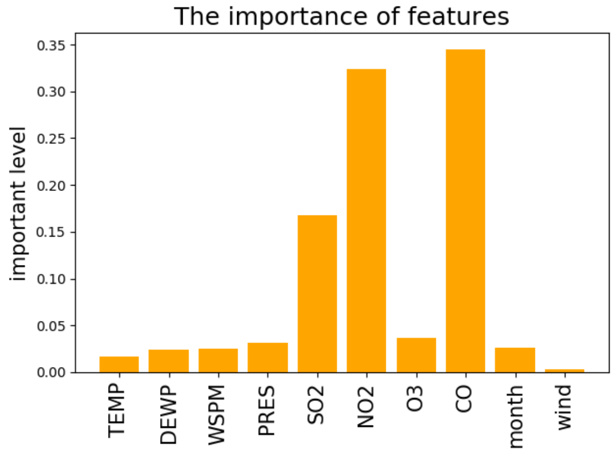


Random Forest

Monthly data

Pollution	Importance
CO	0.345
NO2	0.324
SO2	0.168
O3	0.037

Weather	Importance
PRES	0.031
month	0.026
WSPM	0.025
DEWP	0.024
TEMP	0.017
wind	0.003



Pros and Cons

Pros:
- Reduced error
- Handling of huge amount of data
- No problem of overfitting
- Useful to extract feature importance

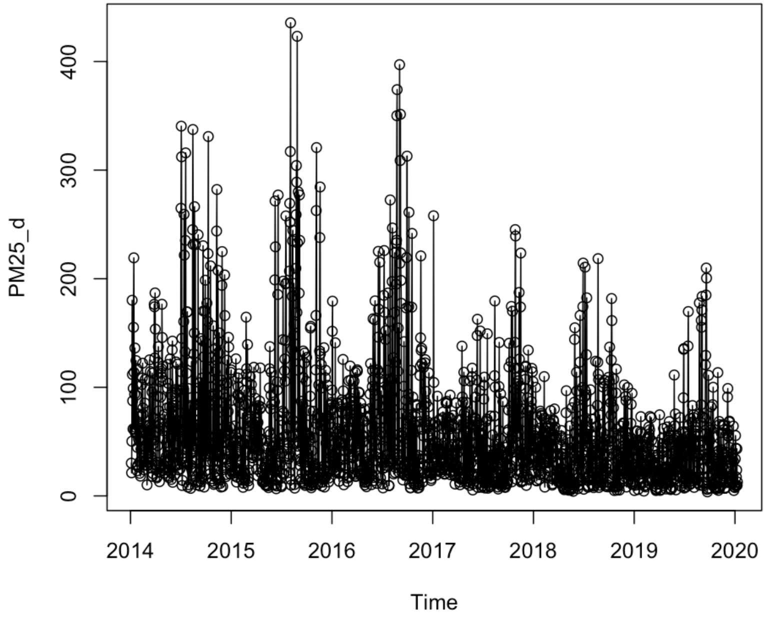
Cons:
- Features need to have some predictive power else they won’t work
- Predictions of the trees need to be uncorrelated
- Appears as Black Box

Time Series Model

- ADF test
- ACF & PACF
- ARMA model
- ARIMA model

Time Series --- Daily

Daily Observations:



Time Series --- Daily

Augmented Dickey-Fuller (ADF) test:

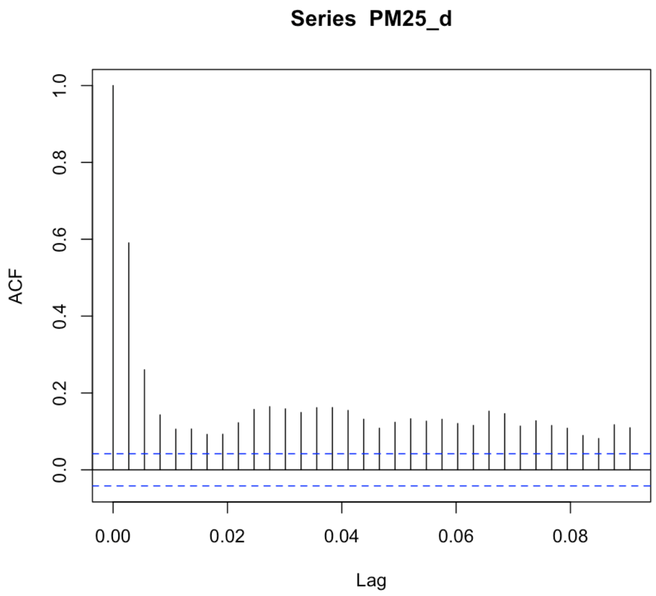
H0: If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary.

H1: The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary.

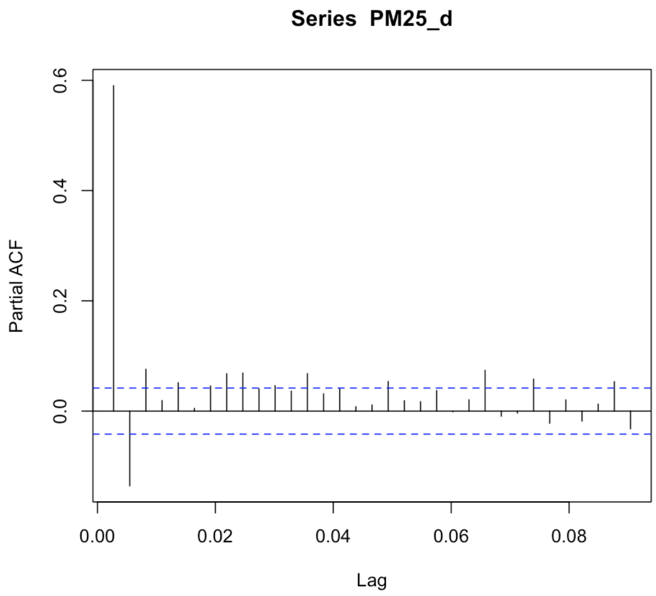
	Value
Test Statistic Value (ADF)	-9.4194
P-value	0.01
Lags Used	12
Critical Value(1%)	-3.4332
Critical Value(5%)	-2.8628
Critical Value(10%)	-2.56744

Note: p-value smaller than printed p-value

Time Series --- Daily



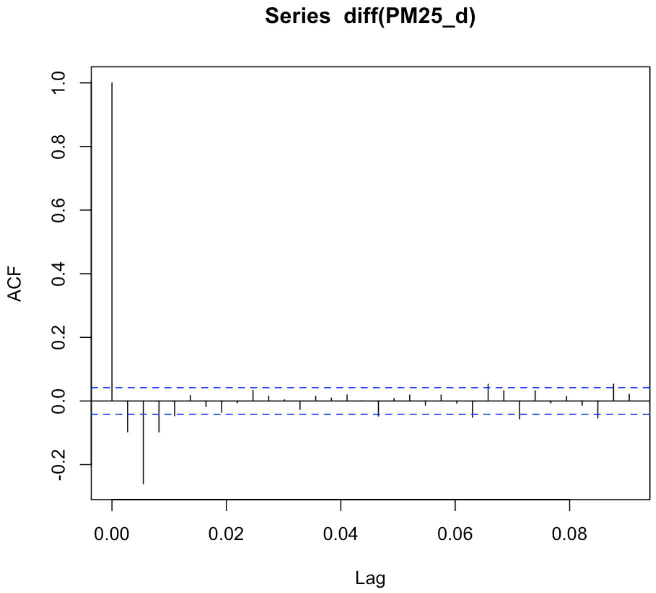
ACF: Tails off



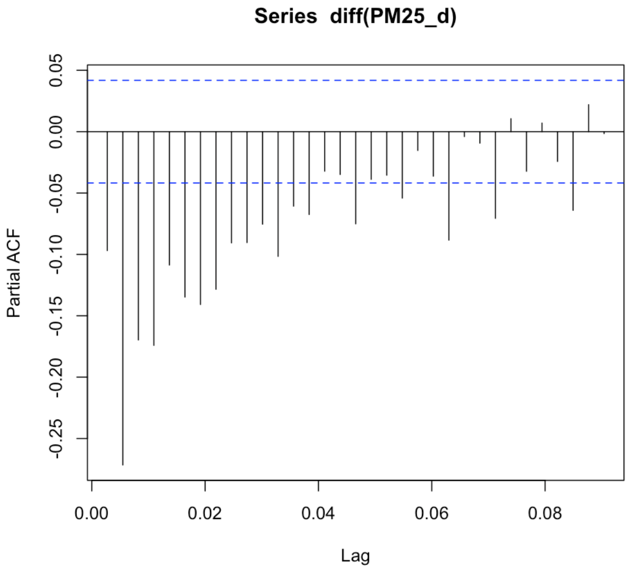
PACF: Cuts off after lag 2

Model: ARMA(2,0)

Time Series --- Daily



ACF: Cuts off after lag 1 or 2 or 3 or 4



PACF: Tails off

Model: ARIMA(0,1,1) or ARIMA(0,1,2) or ARIMA(0,1,3) or ARIMA(0,1,4)

Time Series --- Daily

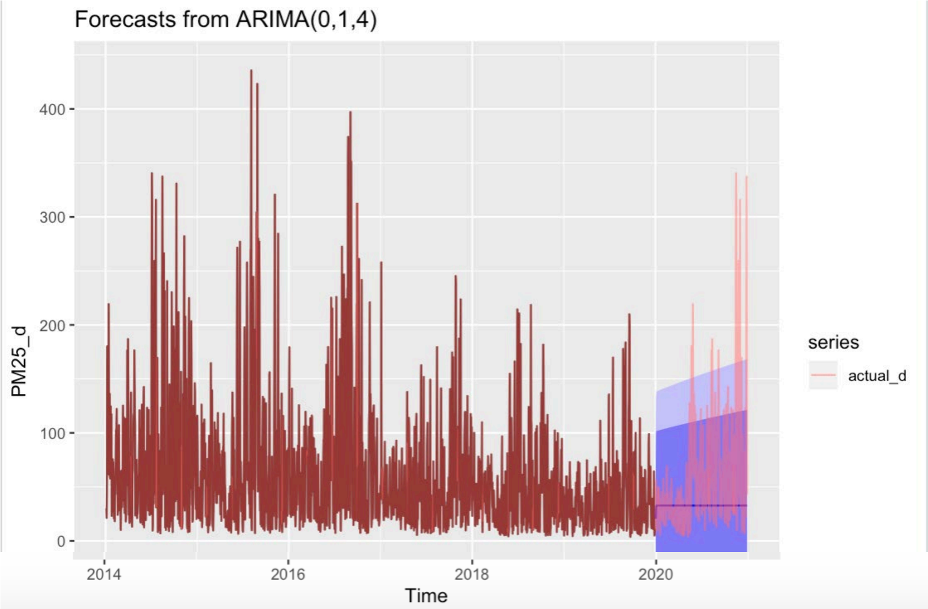
Comparing Models:

Models	AIC
ARMA(2,0)	22837.77
ARIMA(0,1,4)	22767.33
ARIMA(0,1,3)	22771.33
ARIMA(0,1,2)	22820.72
ARIMA(0,1,1)	23313.84
ARIMA(4,1,3)	22770.11

Note: ARIMA(4,1,3) is calculated by 'auto.arima(PM25_d)'

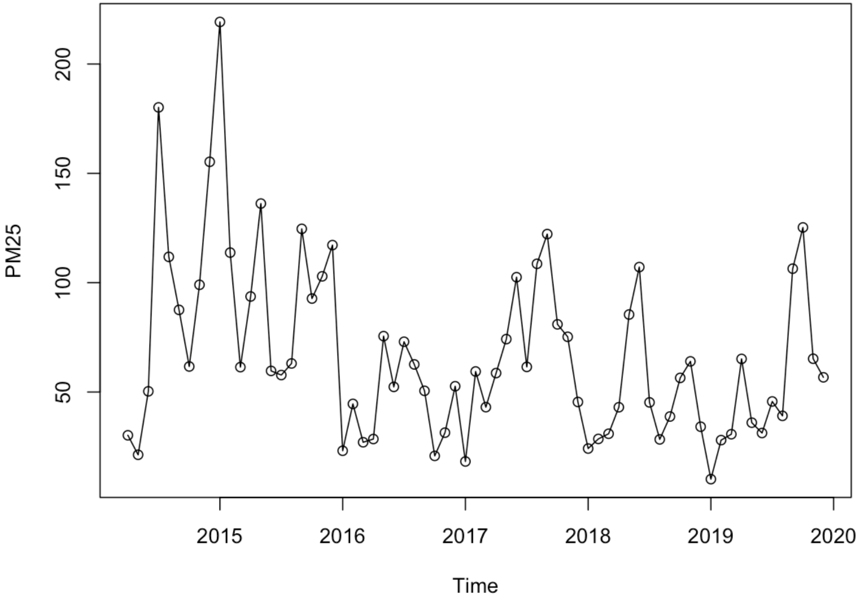
Time Series --- Daily

Model: ARIMA(0,1,4)

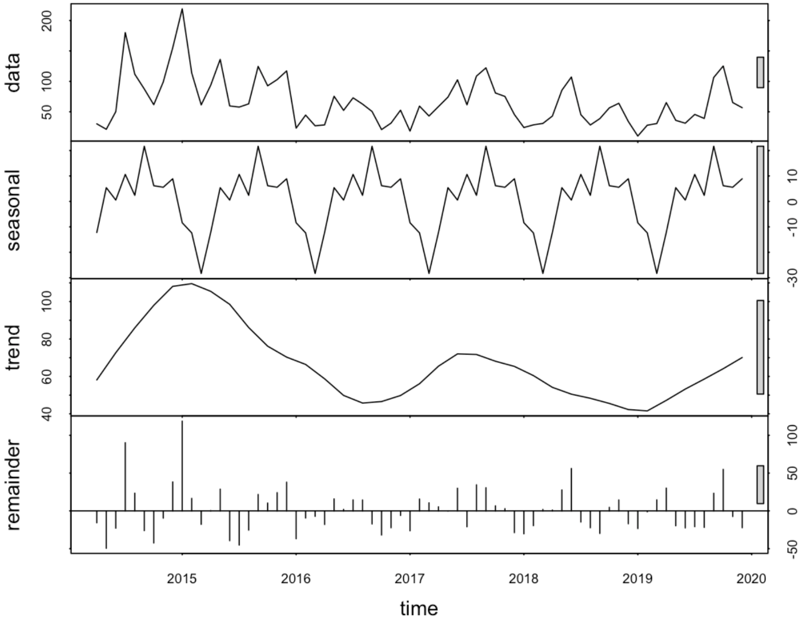


Time Series --- Monthly

Daily Observations:



Check Seasonal Decomposition:



Time Series --- Monthly

ADF test:

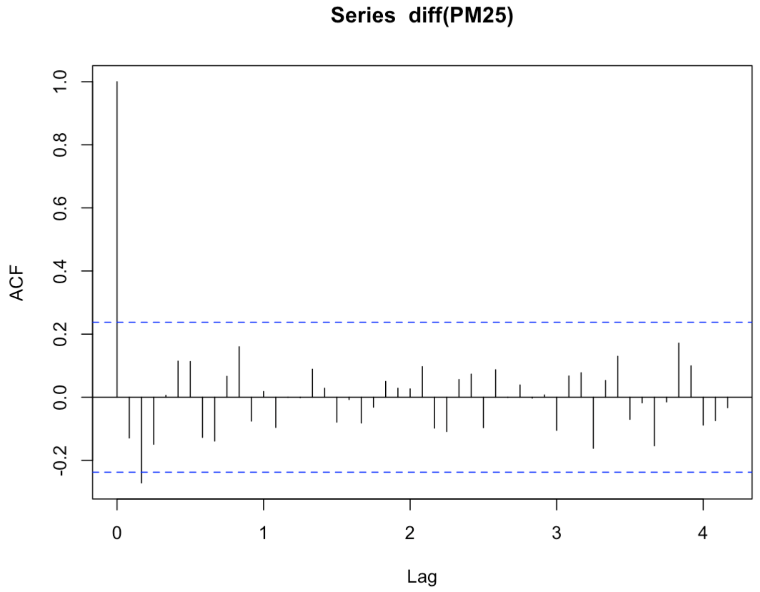
	Value
Test Statistic Value (ADF)	-2.5769
P-value	0.3407
Lags Used	4
Critical Value(1%)	-3.4332
Critical Value(5%)	-2.8628
Critical Value(10%)	-2.56744

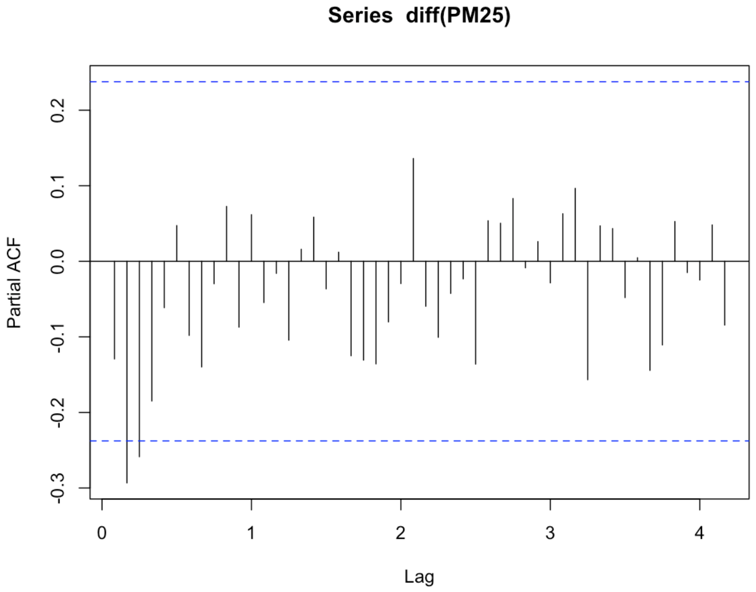
ADF test after diff:

	Value
Test Statistic Value (ADF)	-5.3436
P-value	0.01
Lags Used	4
Critical Value(1%)	-3.4332
Critical Value(5%)	-2.8628
Critical Value(10%)	-2.56744

Note: p-value smaller than printed p-value

Time Series --- Monthly





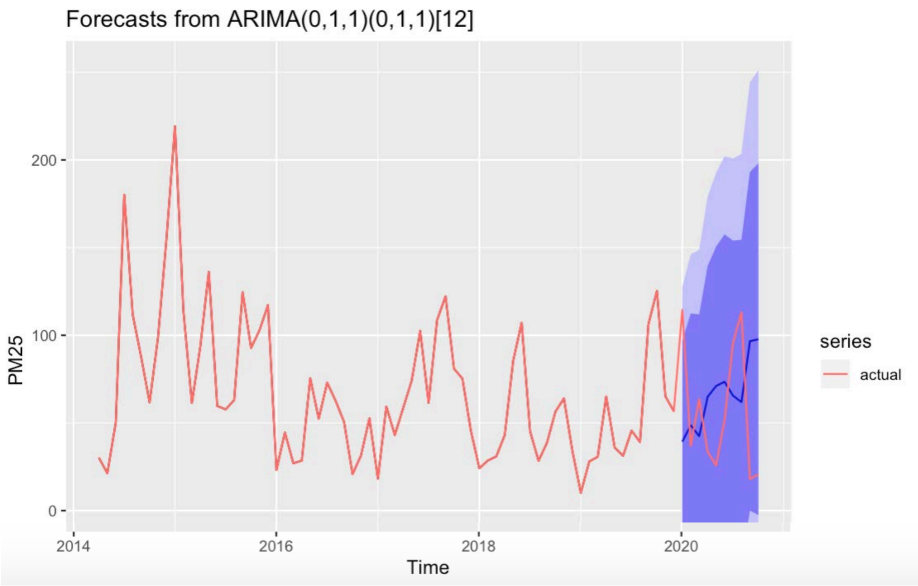
ACF: Non-seasonal: Cuts off after lag 1
Seasonal: Cuts off after lag 1

PACF: Non-seasonal: Tails off
seasonal: Tails off

Model: ARIMA(0,1,1)(0,1,1)[12]

Time Series --- Monthly

Models	AIC
ARIMA(0,1,1)(0,1,1)[12]	596.11

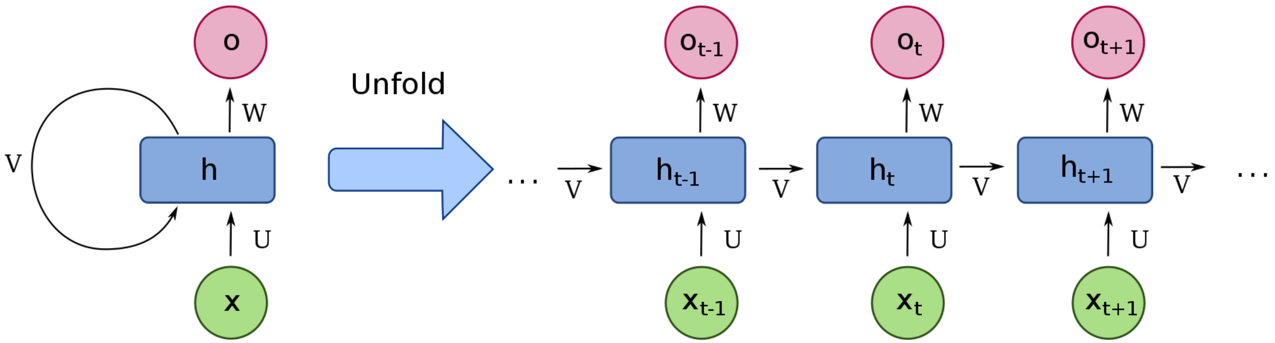


LSTM model

- Basic Structure
- Build the Model
- Predict pm2.5

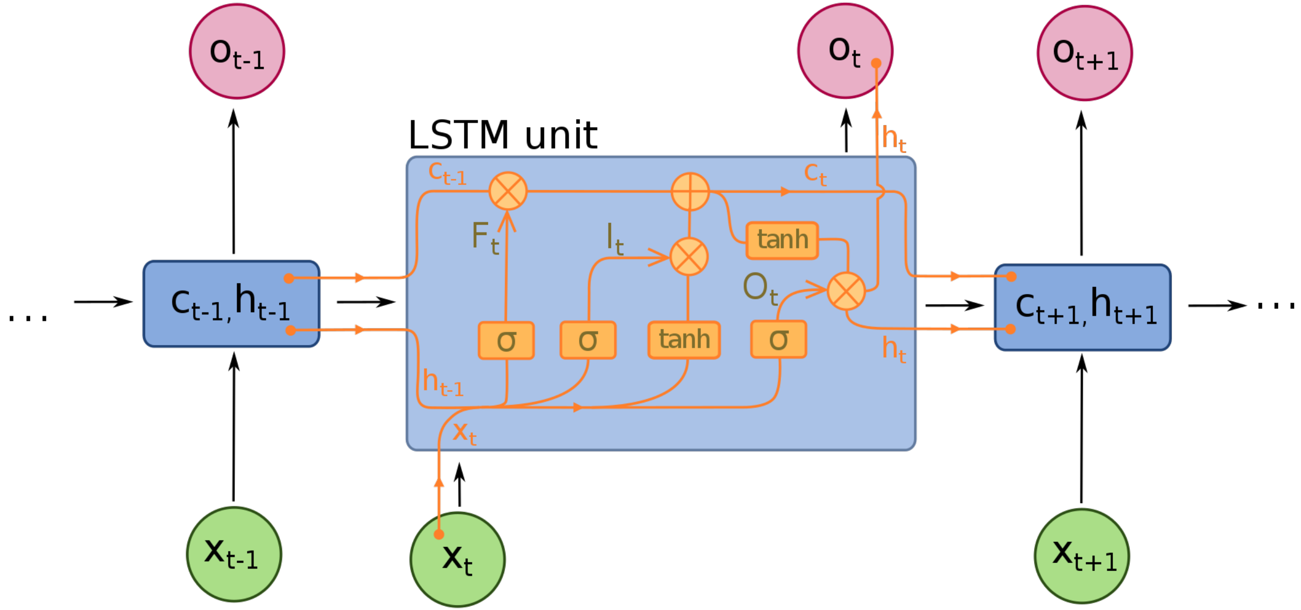
RNN (Recurrent Neural Network)

Basic Structure of RNN model:



LSTM (Long Short Term Memory)

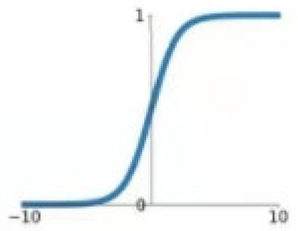
Basic Structure of LSTM model:



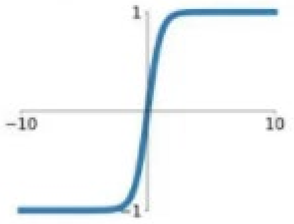
LSTM (Long Short Term Memory)

Activation Function:

sigmoid: $S(x) = \frac{1}{1 + e^{-x}}$



tanh: $tanhx = \frac{sinhx}{coshx} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$



Data Processing

Original data:
- Monthly data: 76*1
- Daily data: 2324*1
- Hourly data: 54385*1

Scaled data:
MinMaxScaler() —— data range (0, 1)

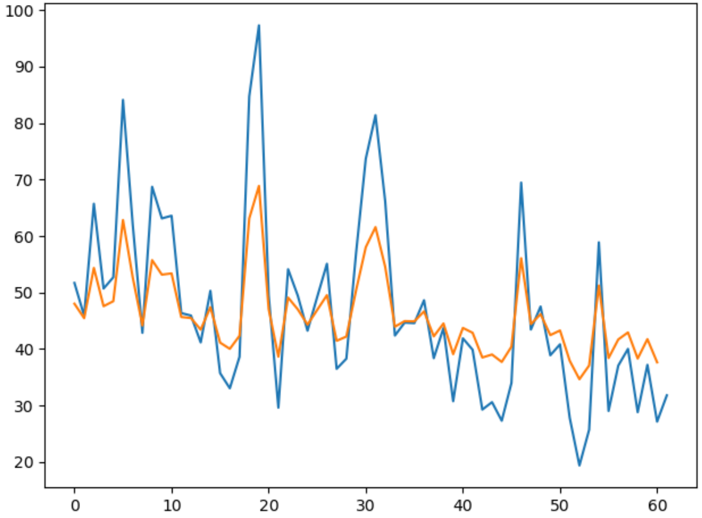
Train set and test set

Build the LSTM Model

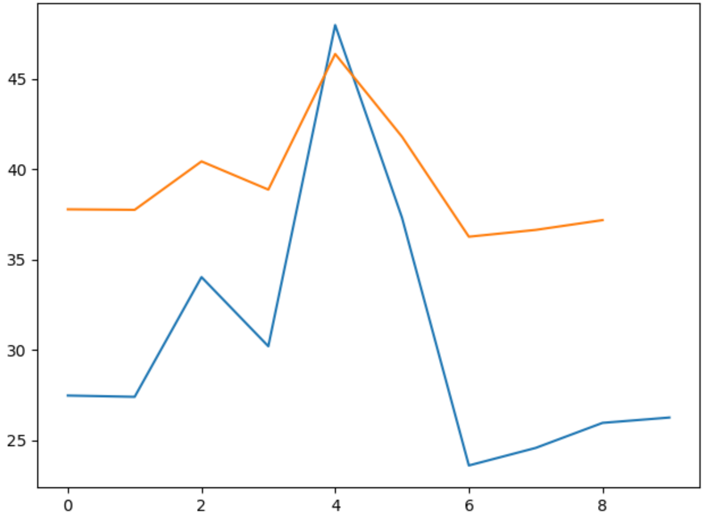
Create and fit the model

hidden layer: 4
loss function:
 mean squared error
optimizer:
 Adam optimizer

LSTM Prediction (monthly data)



train set

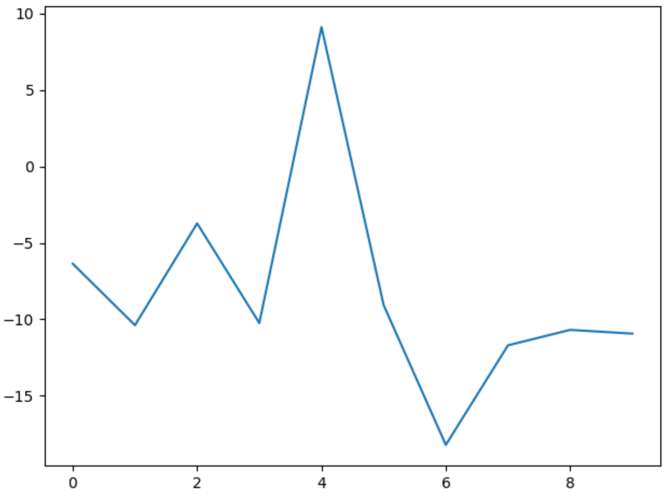


test set

LSTM Prediction (monthly data)

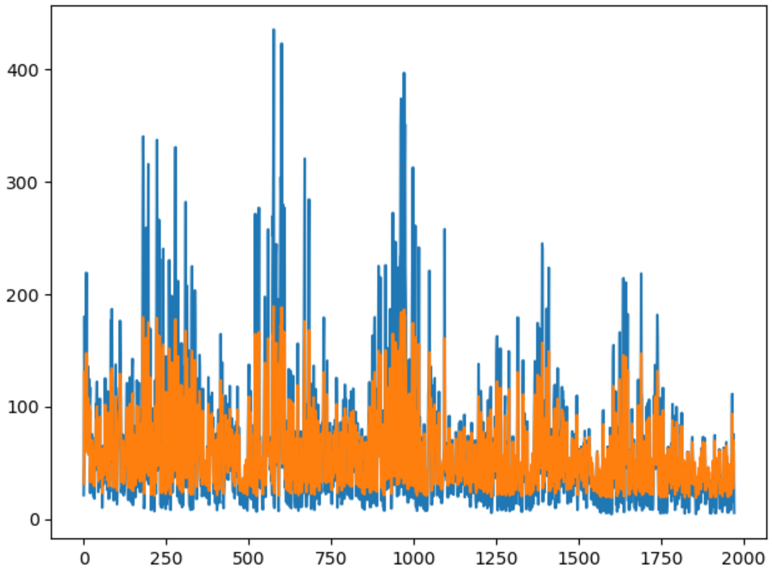
	original	prediction
0	27.466658	33.833000
1	27.395018	37.783016
2	34.028648	37.754608
3	30.196323	40.437862
4	47.991222	38.874893
5	37.316849	46.389771
6	23.592508	41.805149
7	24.564953	36.265778
8	25.951855	36.642929
9	26.247963	37.185135

RMSE:10.6544

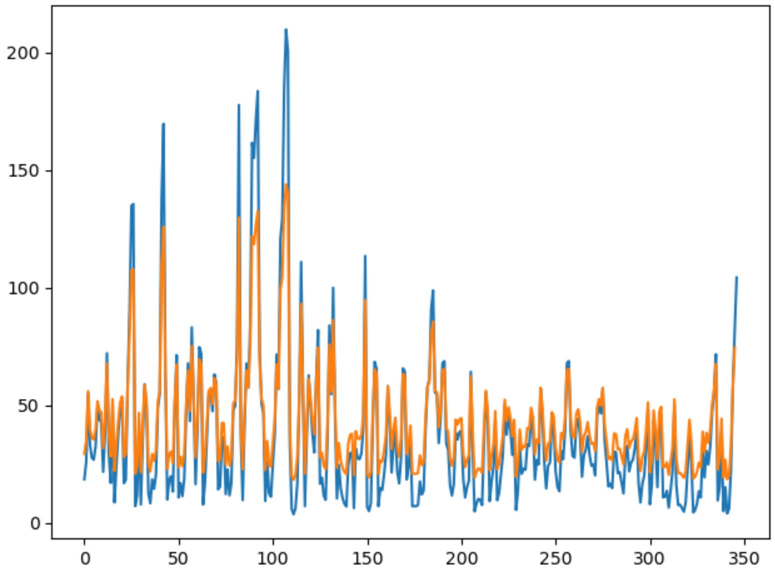


error

LSTM Prediction (daily data)



train set

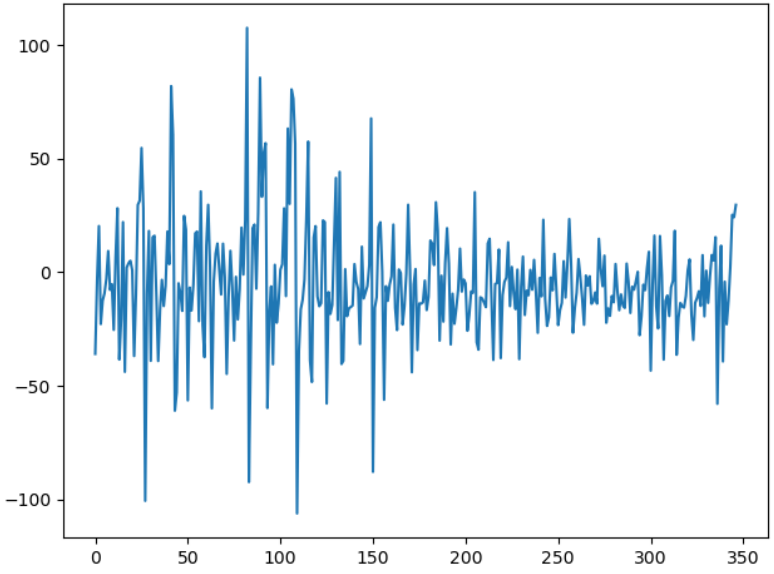


test set

LSTM Prediction (daily data)

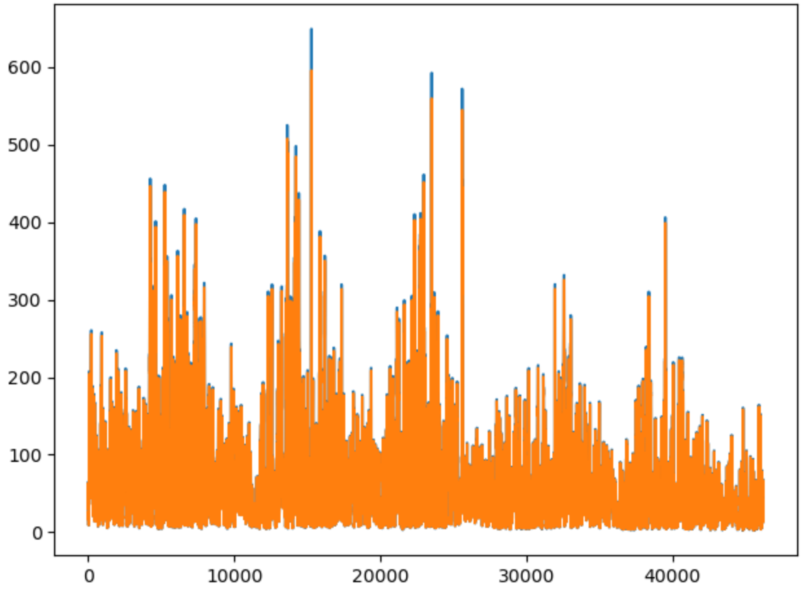
	original	prediction
0	18.621767	54.562695
1	25.961935	29.458778
2	55.270931	34.873837
3	33.288651	56.075130
4	27.775183	40.244186
..	...	...
342	6.104737	18.656055
343	22.715822	20.165522
344	57.863445	32.482944
345	82.196884	57.908207
346	104.413994	74.673843

RMSE:26.3014

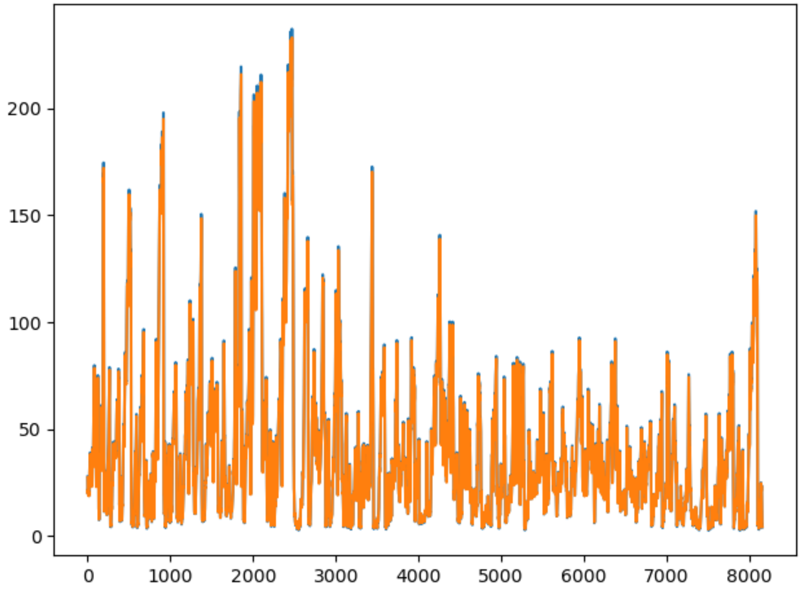


error

LSTM Prediction (hourly data)



train set

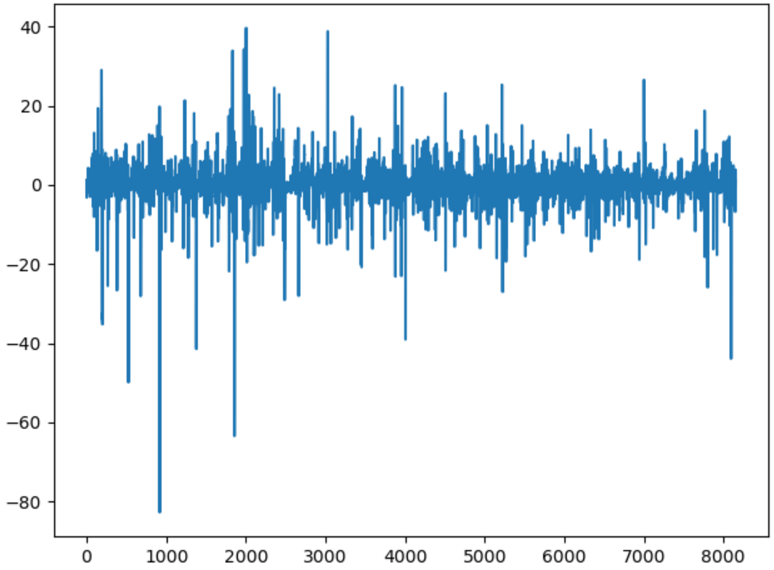


test set

LSTM Prediction (hourly data)

	original	prediction
0	27.806452	26.535864
1	24.580645	27.408976
2	21.090910	24.272615
3	19.909090	20.886841
4	20.781252	19.741976
	...	...
8151	4.833333	10.611366
8152	4.656250	5.223400
8153	4.032258	5.053900
8154	3.937500	4.456697
8155	5.387097	4.366083

RMSE:4.7716



error

Comparison

For LSTM model, the performance of data is different.

Dataset	RMSE
Monthly Data	10.6544
Daily Data	26.3014
Hourly Data	4.7716

Conclusions

- Data preprocessing: missing values, transformation, scaling, merging categorical data level, get different frequency data.

- Regression model:
  Positive relationship: DEWP & pollutants
  Negative relationship: TEMP & month

- Random Forest:
  Pollutants: CO
  Weather: PRES & DEWP

- Time Series:
  ARIMA(0,1,4) daily data
  ARIMA(0,1,1)(0,1,1)[12] monthly data √

- LSTM：
  Hourly data

Reference

- Huang, X.; Yun, H.; Guan, Z.; Li, X.; He, L.; Zhang, Y.; Hu, M. Source apportionment and secondary organic aerosol estimation of PM2.5 in an urban atmosphere in China. Sci. Sin. Terrae 2014, 44, 723–734.
- Ye, W. Study on source apportionment of PM10 and PM2.5 in ambient air of Ningbo. Res. Environ. Sci. 2011, 33, 66–69.
- Shi, J.; Yuan, D.; Zhao, Z. Residential indoor PM2.5 sources, concentration and influencing factors in China. J. Environ. Health 2015, 32, 825–829.
- Beijing Municipal Government. 2016 [cited 2017-02-16]. Available from: http://www.bjstats.gov.cn/tjsj/tjgb/ndgb/201702/t20170227369467.html.
- Liu Z, Hu B, Wang L, Wu F, Gao W, Wang Y. Seasonal and diurnal variation in particulate matter (PM_{10} and $PM_{2.5}$) at an urban site of Beijing: analyses from a 9-year study. Environmental Science & Pollution Research. 2015;22(1):627–642.
- Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., . . . Chen, S. X. (2015). Assessing Beijings PM 2.5 pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 471*(2182), 20150257. doi:10.1098/rspa.2015.0257
- Kai Guo. (n.d.). Beijing releases list of top air polluting sources. Retrieved from http://www.chinadaily.com.cn/a/201805/15/WS5afa22eca3103f6866ee8561.html
- Yuan, G., & Yang, W. (2019). Evaluating China's air pollution control policy with extended AQI indicator system: Example of the Beijing-Tianjin-Hebei region. Sustainability, 11(3), 939.
- Liu, Z., Chen, X., Cai, J., Baležentis, T., & Li, Y. (2020). The Impact of "Coal to Gas" Policy on Air Quality: Evidence from Beijing, China. Energies, 13(15), 3876.