5002 Final Report
Air Quality Prediction

Name : WANG Ruolan

Stu.No : 20551328

ITSC : rwangbh

Content

1. Introduction

Accompanying the rapid urbanization, many developing countries are suffering from serious air pollution problem. China, as the largest developing country in the world, has the highest deaths rate caused by air pollution. Thus, how to use advanced technological means to alleviate this problem is very important for us technical workers.

Beijing is one of the cities with the most serious air pollution. In this project, the purpose is to predict two days' air quality level by using past more than one year of air quality data and the future weather data. The predicted air quality includes PM2.5, PM10 and O3. The prediction task is very difficult since there are a lot of missing values in the raw data. How to deal with them is one of a crucial point for the accuracy of the prediction result. Also, since the raw data is form different sources, we should firstly integrate them and give them same format. All in all, in order to get a great prediction result, we should be carefully do every step which includes data cleaning, feature engineering and model training so as to get an accurate predicting result.

2. Data preprocessing

2.1 Data Integration

Since the raw data is collected from different resources, they have different attribute name and data format. I firstly change these attribute name with a uniformed name and unified same type data which have different time span into one dataset. For example, I change the name PM25_concentration in the airqQuality_201804 to PM2.5, which is as the same as other files.

2.2 Data Analysis

After analyzing all given datasets, I found that the air quality data and grid weather station data are directly related to the prediction of AQI (Air Quality Index). And the observed weather data is not used in this project. In order to remove the effect of noise, I used the data of month 3, 4, 5, 6 in 2017 and 2018 as the raw dataset, which I based on to do feature engineering work and the final prediction.

The concrete information about the integrated air quality data and grid weather data are shown in the table below.

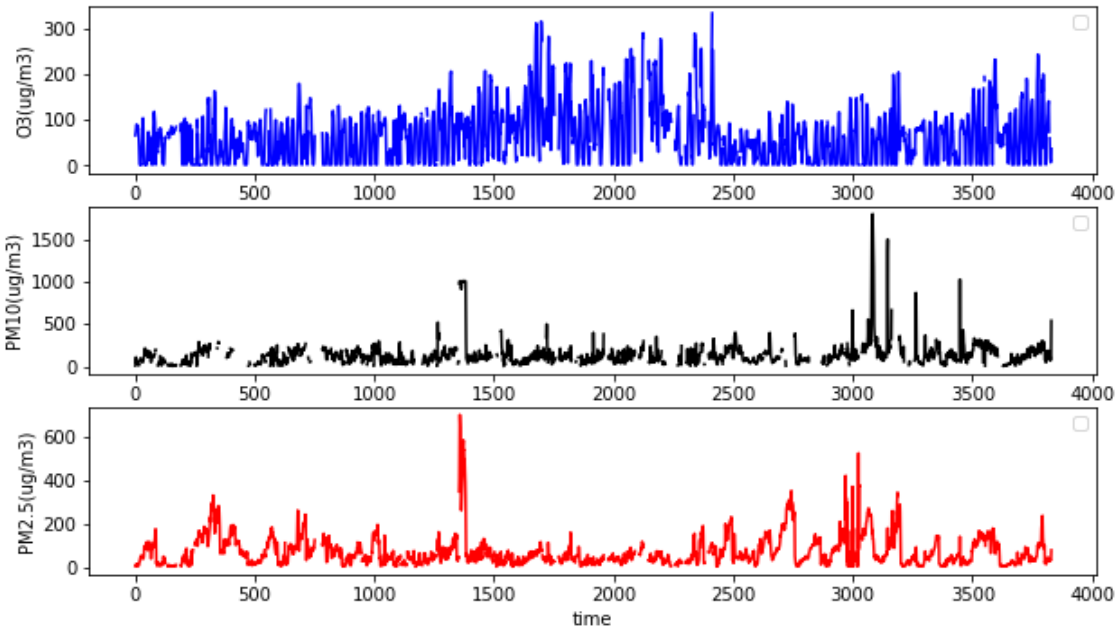
Attributes	missing rate
PM2.5	5.75%
PM10	23.71%
O3	5.31%
NO2	4.96%
CO	14.39%
SO2	4.67%

Table 1 Air Quality Information

Attributes	miss rate
humidity	0
pressure	0
temperature	0
weather	82.40%(no data in 2017)
wind_direction	0
wind_speed	0

Table 2 Grid Weather Information

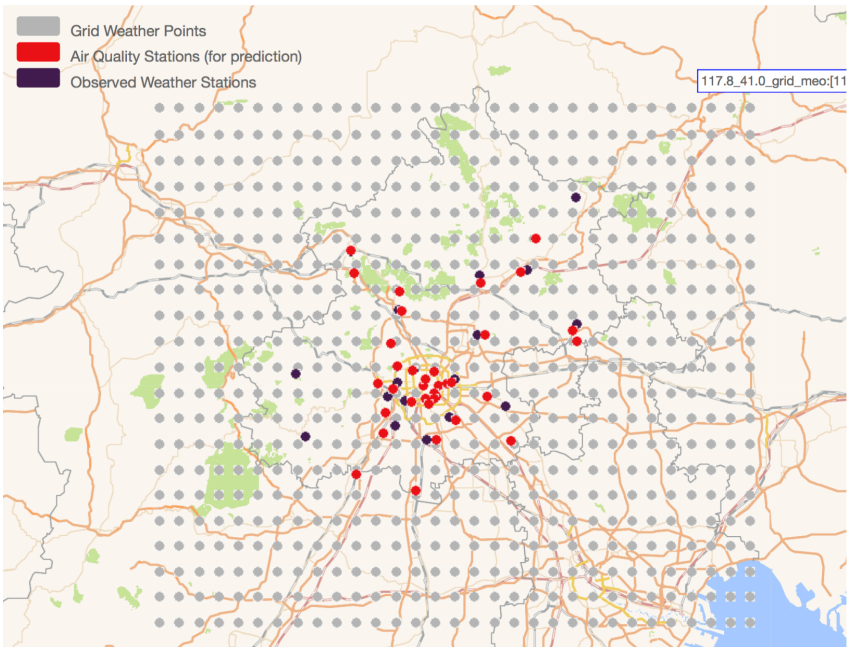
After analyzing the tend of PM10, PM2.5 and O3, I find that these attributes don't has obvious tendency and periodic. The picture below shows the trend of these three pollutants in two months, from 2018-3-1 to 2018-4-30.



Picture 1 Pollutant Change

As is shown in above picture, PM10 has many missing values, which means we should at first predict the label in order to have a further analysis. In PM10 and PM2.5, there are some peaks which are highly different from the values of ordinary days. This gives us a reminder that some particular days like vacations and weekends should be paid a lot of attention when predicting.

The relative spatial location of grid weather stations and air quality weather stations is shown in picture 2. Since we cannot directly get weather, humidity and pressure information of the air quality station, we should use its special relationship to get information of these stations indirectly.



Picture 2 Station Location

2.3 Missing Value Interpolation

As is shown in table 1 and table 2, the attribute PM10 has the largest missing ratio in the air quality table. And other attributes also have different levels of missing value. In grid weather data, it shows that the weather is missing in all the months in 2017. Thus, for different missing value level and different conditions, I use different methods to deal with them in order to get a better prediction.

(1) Spatial interpolation

In the attributes of PM10 and weather, the circumstance that data is missing during a long consecutive time happens. We cannot predict the missing value by using polynomial interpolation since the values near the missing value is also missing. In this condition, if we use interpolate method to fill in, it will bring a lot of noise. Then I used a classical spatial interpolation method, inverse distance weighting (IDW), to interpolate this kind of missing value. I filled missing value by using the information of the nearest three stations. The larger distance to the missing value station, the smaller the influence it will have to the missing value station. It follows the First Law of Geography, " Everything is related to everything else, but near things are more related than distant things".

(2) Arima prediction

However, the spatial interpolation cannot solve the problem that all nearby values are missing when predicting the missing values. Thus, I use the character of time series to fill in some left parts of missing values.

As we all know, Arima model is one of the most effective methods to deal with time series problem because the model usually does well in capturing the change pattern of time series data. Thus, I use this model to predict the left part of missing value of air quality data. I use the data before the missing value to predict the missing one. And then, I recurrently do this step to predict other missing values.

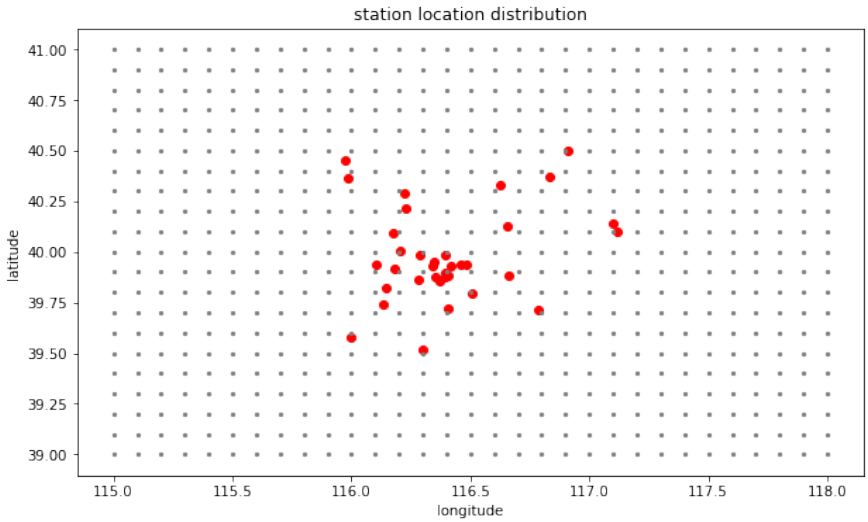
(3) Backward Filling

For the data which are missing during long time span, using Arima model to predict would take a lot of time. Thus, I used backward filling to fill in the missing values.

3. Feature Engineering

3.1 Spatial feature construction

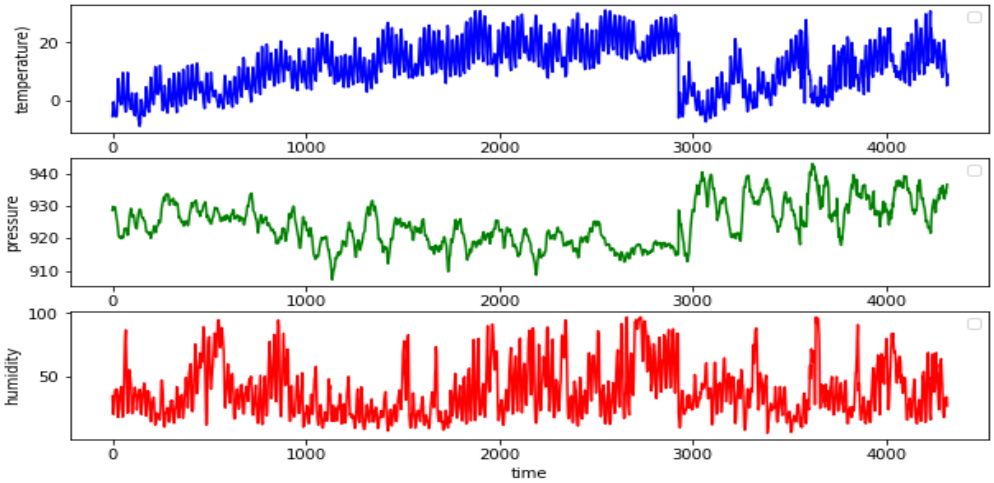
As is shown in Table 2, we have the information of humidity, pressure, weather and temperature for every grid weather station. However, we cannot obtain that information directly for air quality stations. The relative location of grid weather stations and air quality stations can be shown in the picture below. Thus, based on the location information, I use the average information of grid weathers which are in the k-radios of the target air quality station to represent the humidity, pressure, weather and temperature information.



Picture 3 Station Distribution

3.2 Domain Feature construction

As the picture shows below, humidity, pressure and temperature changes day by day. After analyzing the data, I construct difference features for humidity, pressure and temperature as domain feature. I use the difference of humidity, pressure and temperature between one hour the last hour as new features. In this way, I can capture time series information by making last hour's information as features. The change of humidity, pressure and temperature can be shown as picture 4.



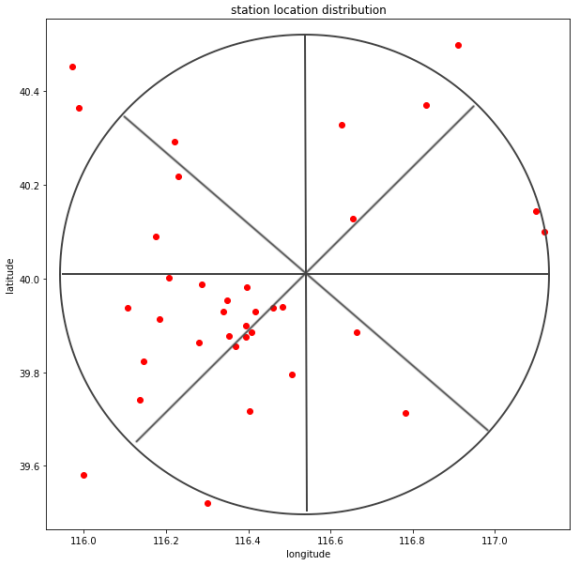
Picture 4 Grid Station Information

3.3 Statistical feature construction

For every station, I construct statistical features of O3, PM2.5 and PM10. These features include standard deviation, the maximum and minimum value of these three air quality data. For different stations, the change ratio of air quality is very different.

3.4 Wind direction

When constructing wind feature, I consider to decomposing the wind speed into x-axis and y-axis. For wind direction, I normalize it into eight directions, which are east, west, south, north, northwest, northeast, southeast and southwest. The direction can be shown in the picture below.

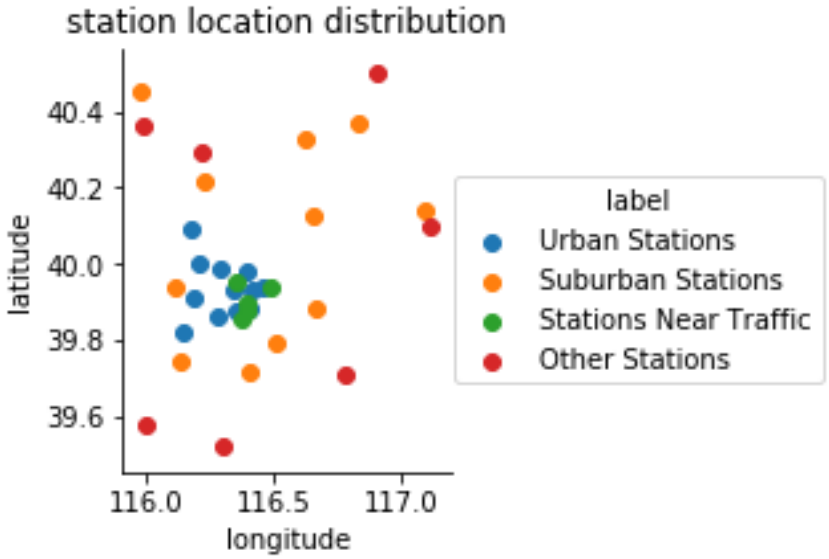


Picture 5 Wind Direction

3.5 Location Feature Construction

The location of the air quality station has been given, which are Urban Stations, Suburban Stations, Stations Near Traffic and Other Stations. The station location is a very important feature for air quality prediction. If the station is located near the

traffic, which is usually crowed of people and cars, the air quality would be very terrible. On the country, if the station is located in the suburban, the air quality would be better than others.



Picture 6 Station Label

4 Model Construction

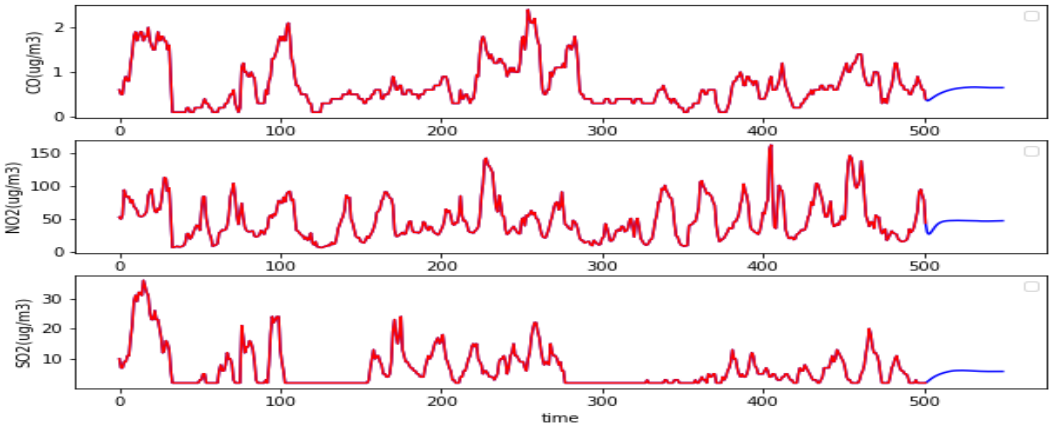
4.1 Model Evaluation

In this project, I use smape to evaluate the accuracy of the model. The calculation formula can be shown as following.

$$smape = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(A_t + F_t)/2}$$

4.2 Time Series Model

After analyzing, I found that CO, SO2 and NO2 are very important features. However, we don't have that information during 2018-05-01 and 2018-05-02. Thus, I try to use Arima model to predict the information of CO, SO2 and NO2 step by step. I use the past 1,000 step data to predict the next hour's air quality data iteratively. The below picture shows the prediction result.

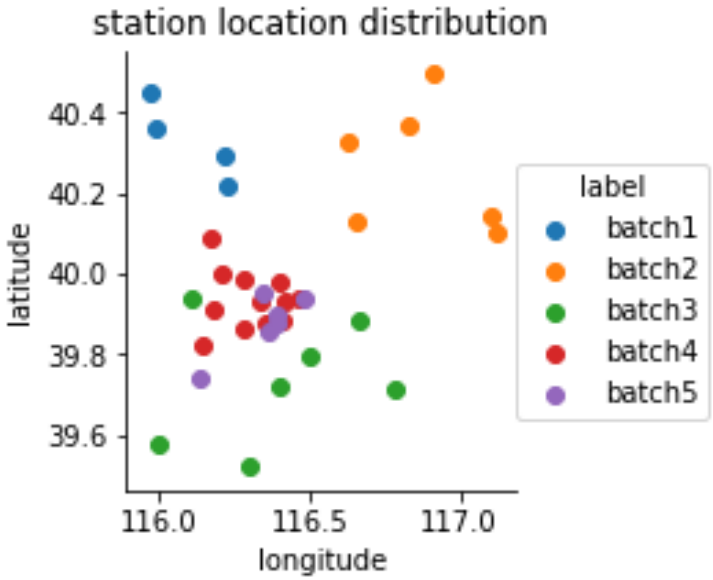


Picture 7 Arima Model Prediction

As it is shown in above picture, I found that in the condition of predicting 48 steps forward, Arima model cannot capture the change pattern of the data. Thus, it is not a good method to use it to predict the air quality data and fill in the missing values which miss during long time. As is shown in the above result of Arima model, time series model performs bad for this problem since the prediction result will become worse and worse with the steps increases. Thus, I didn't use it for further prediction.

4.3 Tree-based model

I use Lightgbm to predict the AQI. By combining the station type and spatial location information of stations, I divided them into five batches. Since Stations Near Traffic and Urban stations are gathered in one place individually. I build models for them respectively. The Suburban Stations and Others Stations are very sparsely, I construct models based on their location information. For the stations which are near each other, I put them into one batch. The concrete information can be shown as the picture below.



Picture 8 Station Batches Location

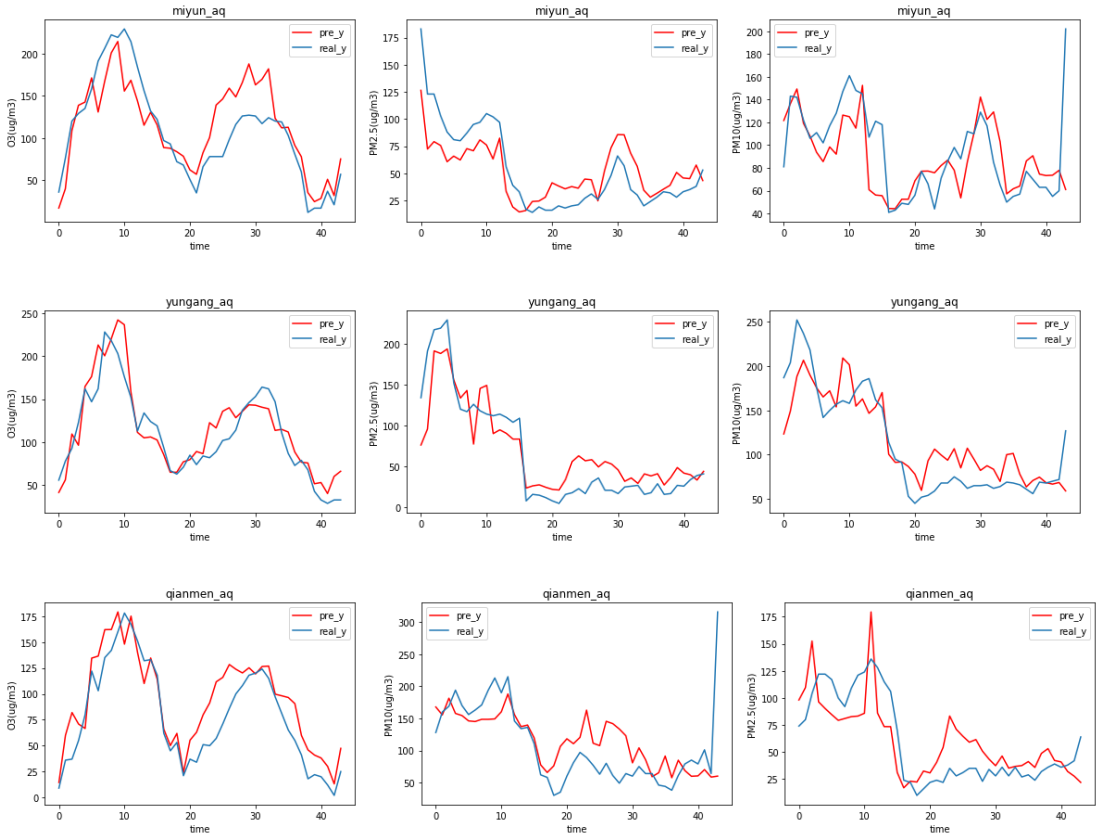
For different batches of stations, I use 20% data as validation set and use the data from 2018-04-28 to 2018-04-30 as testing set to define the accuracy the trained model. The other data will be used as training set to train the model. I used lightgbm to train different models for different batches. Finally, I calculated the smape to evaluate the prediction results. And the smape of the testing data is showed as the table below.

Batch	O3	PM2.5	PM10
Batch1	0.393	0.406	0.398
Batch2	0.253	0.410	0.278
Batch3	0.317	0.392	0.340
Batch4	0.319	0.398	0.335
Batch5	0.377	0.450	0.323

Table 3 Training Result

As is shown in picture 8, batch1 has four stations, batch2 has six stations, batch3 has seven stations, batch4 has twelve stations and batch5 has six stations. By averaging the above training result, the testing smape of O3 is 0.325, the PM2.5 is 0.408, the PM10 is 0.331.

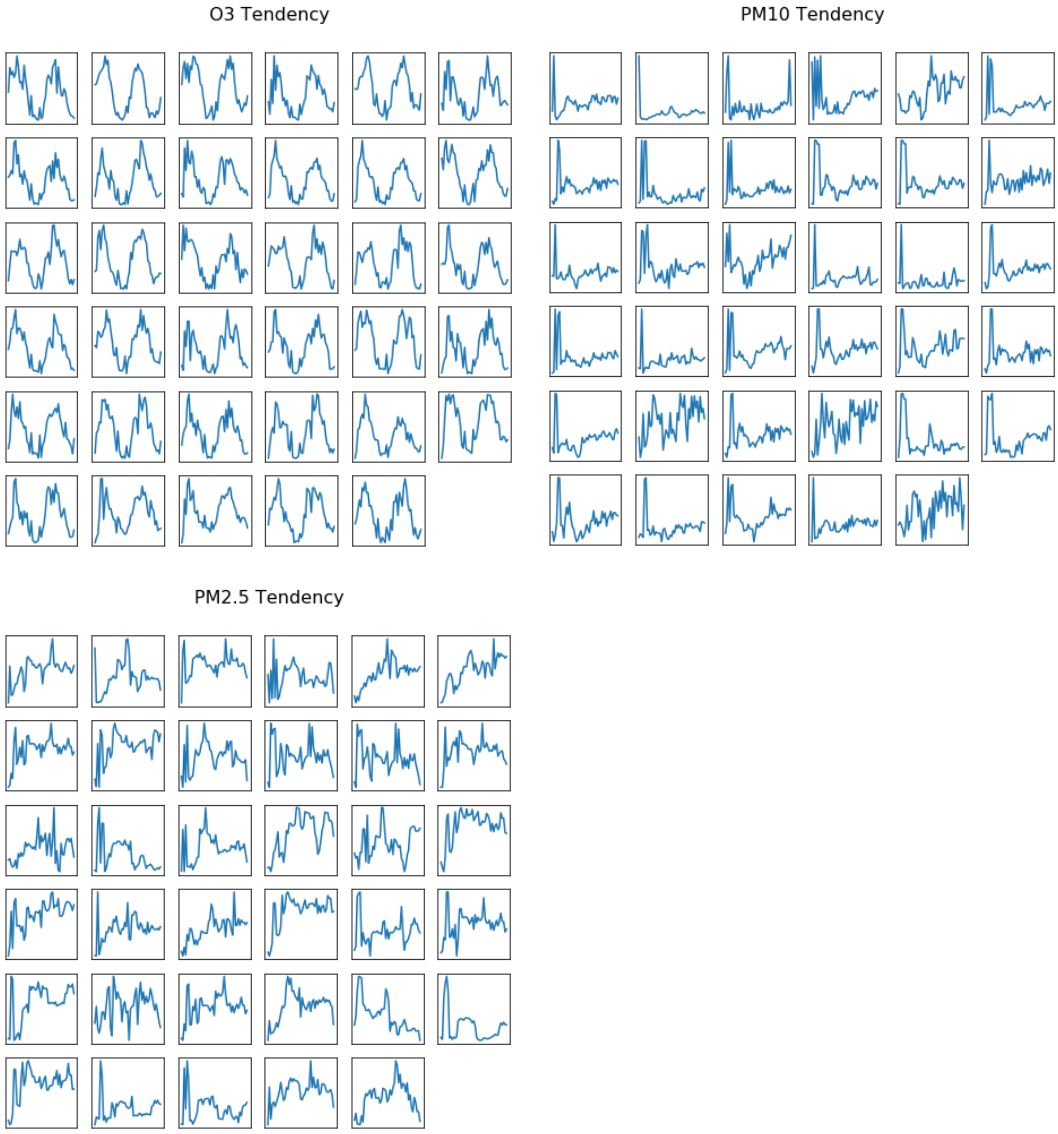
I selected several stations to show the final result here.



Picture 9 Prediction Results

4.4 Final Prediction Results

For the prediction of O3, PM10 and PM2.5, I used 20% data as validation and the others as training data. The AQI tendency of 35 stations are shown as follows.



Picture 10 Prediction Tendency

As the picture shows above, the tendency of O3 for 35 stations are very similar from 2108-5-1 to 2018-5-2 and it changes very regularly. However, the level of PM2.5 and PM10 change very irregularly during the two days, which means they changes unstably.

5 Advantages and Disadvantages

In this project, I considered very comprehensively when dealing with missing values and constructing features. For different type of missing value, I used different methods to fill in. I construct features from five aspects by combining the spatial the temporal information. In the model training part, I divided the stations into five parts

to train, which can save a lot of time and improve accuracy since the amount of training data is larger compared with training model for every station.

However, in the model construction part, I only use Lightgbm to predict the final result because of the limitation of time. And I didn't use ensemble method to improve the accuracy of the model. In the prediction of air quality feature, I didn't find a great method to predict accurately, which may cause some noise for the final result.