

A Proposal to Analysis and Visualization of China National Air Quality Dataset of March 2017

Yifang Cao

Introduction

China is now the world's second largest and fastest-growing major economy with an average growth rate of 10% for the past 30 years. However, its fast industrialization and development rate persists at the cost of the environment. There are now three major crises that are confronting China, which are resource shortage, environmental pollution and ecology destruction. Among these three, the environmental pollution, especially atmospheric pollution has recently earned its most notorious reputation. To curb the increasing atmospheric pollution, the Executive Meetings of the State Council published the revised *Ambient Air Quality Standard* in February 2012, and continuously introduced many stricter nation-wide policies and regulations regarding to the air pollution in the following years. Because the regional energy consumption structure and productivity may vary over time, the geographical pattern air pollution may also vary. Thus, a latest analysis on the most recent air pollution pattern is necessary. The analytical results may help the authorities adjust the current regulations or introduce new targeted regulations in time.

This document is intended to propose a project that analyzes and visualizes China national air quality dataset of March 2017, which is the most recent data as of the time of proposing. This proposal will discuss the needs and the urgency of this project, the scope of work, and the objectives and approaches that will be used to conduct the project.

Statement of Needs

Air pollution is no stranger to the world. In December 1952, the most infamous air pollution incident occurred in London. The airborne pollutant caused mainly by coal consumption directly and indirectly led to 12,000 fatalities less than a week. This incident has shown how hazardous and acute air pollution can be without proper regulations. Many medical studies have shown that the direct consequences of air pollution are respiratory and cardiovascular diseases. Among all the airborne pollutant, particulate matter that contains heavy metal chemicals can remain in alveoli and will cause pulmonary sclerosis, and eventually lung cancer.

China, as a fast-developing country, is now going through the same air pollution problems. The statistics published by the National Health and Family Planning Commission of the PRC indicates that the lung cancer rate has been growing 26.9% every year, and mortality caused by lung cancer has risen 465% over the past 30 years. The high cancer rate is a factual indication of the hazardous air quality in China. Since 2012, authorities have established laws and taken effective and strict measures on air pollution controls. The efforts have led to preliminary results.

Because of the measures taken past years, many high pollution companies were shut down or reformed and new companies opened, the regional industrial structure may not remain the same. To ensure that the effectiveness and accuracy of the current air pollution regulations, analysis on the latest geographical features of air quality is needed. The proper interactive visualization is necessary to convey the results promisingly and compellingly.

Scope of Work

The air quality dataset used in this project are drawn from SinaApp, a website that collects and organizes national air quality historical data originating from China National Environmental Monitoring Center's Hourly Air Quality Index(AQI) Monitoring Data. AQI observations are collected hourly from 1497 ground monitoring stations all over China. The dataset is organized in a way that each type of observation has one row for each hour being recorded, like the following example:

date	hour	type	1001A	...
20170301	0	AQI	27	...
20170301	0	PM2.5	15	...
20170301	0	PM2.5_24h	10	...
20170301	0	PM10	27	...
...

The first row contains all the attributes. The attribute “1001A” in the example means the station code 1001A. There are 1497 such station code attributes in all that contain column data, and 12 different observation types. The first row, for instance, means on March 1 2017, the AQI value for the environmental observation station 1001A is 27. There is no unit provided in the original dataset. However, SinaApp provides the following information about the unit and explanation of each data type:

Type	Data Type	Unit
PM2.5	Particulate Matter 2.5-micron diameter	μg/m3
PM2.5_24h	Average PM2.5 for the past 24 hours	μg/m3

PM10	Particulate Matter 10-micron diameter	μg/m3
PM10_24h	Average PM10 for the past 24 hours	μg/m3
AQI	Air quality index for the station and hour	N/A
SO2	Sodium dioxide	μg/m3
SO2_24h	Average sodium dioxide value for the past 24 hours	μg/m3
NO2	Nitrogen dioxide	μg/m3
NO2_24h	Average nitrogen dioxide value for the past 24 hours	μg/m3
O3	Ozone	μg/m3
CO	Carbon monoxide	mg/m3
CO_24h	Average carbon monoxide value for the past 24 hours	mg/m3

Source: Beijingair.SinaApp

The analysis and visualization in this project will only focus on data obtained from March 1 2017 to March 31 2017. There are two reasons for this specific choice to estimate the latest China national air quality. The first is simply that this is the latest dataset that can be obtained as the time of proposing this project, which can best reflect and satisfy the requirement of analyzing the most recent air quality condition. The second reason is that the massive national concentrate heating period ended on March 15 2017 in the most area of China. As many statistics that has shown, air quality is usually noticeably worse during the winter, because of the high fossil fuel consumption for heating. The first half of March is an intensive period of high fossil fuel consumption, and the second half is not. Thus, the dataset of March can be used to yield a good estimation of the average annual air quality.

In addition, for better interactive visualization, another dataset containing location information on 1497 monitoring stations will be also in this project, which will help visualize geographic features of the recent pollution pattern. The dataset is published by SinaApp, which is organized as following:

Station Code	Station Name	City	Longitude	Latitude
1001A	Wanshou West Palace	Beijing	116.366	39.8673
...

The first row means that the environmental monitoring station named Wanshou West Palace with a station code of 1001A is location in Beijing, with the longitude of 116.366° and the latitude of 39.8673°.

Objectives and Approaches

The AQI is commonly used by Chinese government to indicate current air quality and forecast future trends. As the AQI increases, it is more likely that more people will experience increasingly severe adverse health effects. The AQI is mainly depend on 6 types of airborne pollutants, PM2.5, PM10, SO₂, NO₂, O₃, and CO. Therefore, it is important to find how each type of pollutant is correlated with AQI recently. The objectives are following:

1. For each monitoring stations, visualize 6 types of pollutants with bin charts.
2. Visualize values of 6 types of pollutant and AQI with parallel coordinate plot, and search for the correlations.

In order to find geographic features of the air pollution pattern, the following analysis and visualization will also be conducted:

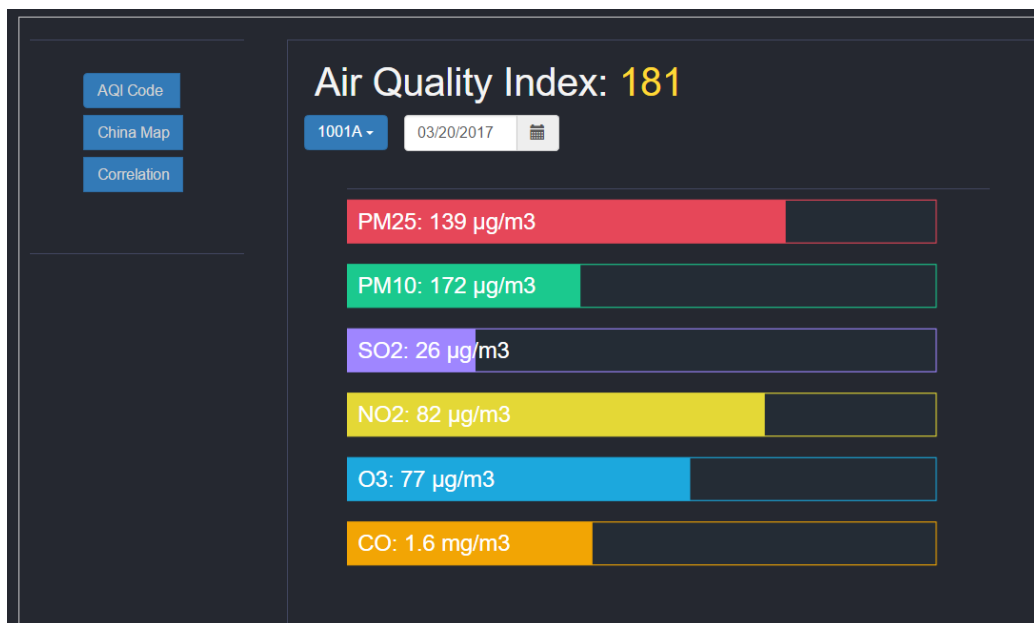
3. Visualize the geographic distribution of each type of air pollutant with choropleth map, which will support brushing feature.
4. Find the pollution centers and clusters for each type of airborne pollutant.

Progress Report

Interface

To provide an interactive interface to users, a client-server model was implemented. The server was implemented with Python Flask along with MongoDB as the back-end data store to store the air quality data provided by China National Environmental Monitoring Center. Some Python mathematics libraries such as sklearn and numpy were used to perform data analysis. In the front end, JQuery and Bootstrap were used to facilitate development and provide interactive layout. D3.js was used to help data visualization, and Anymap.js was used to help visualize pollution patterns on a map of China.

The following image shows how the basic dashboard was laid out. The left side of the page is the navigation menu, which consists of three buttons, AQI Code, China Map, and Correlation. The button AQI Code will lead to a page that visualizes 6 types of pollutants with bin charts for each monitoring station. The China Map button leads to a page that visualize pollution patterns of each type of pollutants. The Correlation button leads to a page that display the choropleth map.

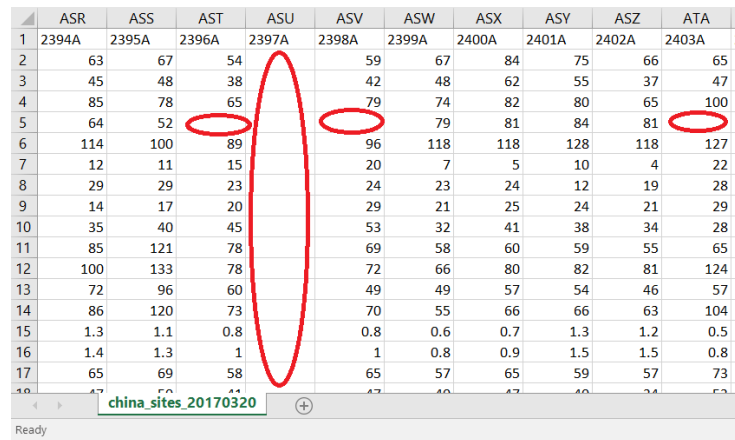


The right side of the page is the main displaying area for each type of data visualization.

Data Preprocess

The data files provided by SinaApp were in CVS format, and one file for each data. Because data of each date is large, the data analysis and visualization will be performed separately for each date. Each file for each date was stored into one MongoDB database but into separate collections. The data file that contains the information of monitoring station locations was stored into a separate database.

In all the data files provided, some monitoring stations are missing all its data records. These monitoring stations usually do not have information in the location data file, which indicates these monitoring stations no long existed. Thus, these monitoring stations were deleted directly.

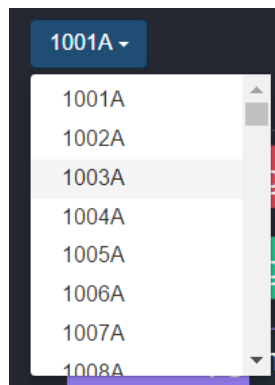


	ASR	ASS	AST	ASU	ASV	ASW	ASX	ASY	ASZ	ATA	
1	2394A	2395A	2396A	2397A	2398A	2399A	2400A	2401A	2402A	2403A	2
2		63	67	54		59	67	84	75	66	65
3		45	48	38		42	48	62	55	37	47
4		85	78	65		79	74	82	80	65	100
5		64	52			79	81	84	81		
6		114	100	89		96	118	118	128	118	127
7		12	11	15		20	7	5	10	4	22
8		29	29	23		24	23	24	12	19	28
9		14	17	20		29	21	25	24	21	29
10		35	40	45		53	32	41	38	34	28
11		85	121	78		69	58	60	59	55	65
12		100	133	78		72	66	80	82	81	124
13		72	96	60		49	49	57	54	46	57
14		86	120	73		70	55	66	66	63	104
15		1.3	1.1	0.8		0.8	0.6	0.7	1.3	1.2	0.5
16		1.4	1.3	1		1	0.8	0.9	1.5	1.5	0.8
17		65	69	58		65	57	65	59	57	73
18		47	50	44		47	40	47	40	34	50

Also, some monitoring stations are occasionally missing some data entries, which may result from incomplete monitoring. Those missing entries should be properly approximated. Because the monitoring stations were encoded per their physical locations, the adjacent monitoring stations in the data file are also very close to each other in the real world. Their recorded values should not vary too much due to proximity. Thus, the neighboring data can be used as a proper approximation for the missing data. For example, in the data file table shown above, the entry (Row 5, Column 2396A) is missing. Its approximation can be obtained either from the entry (Row 5, Column 2395A) or the entry (Row 5, Column 2397A). In this case, the entry (Row 5, Column 2397A) is missing as well, so the former was used. If both adjacent entries are missing as well, the missing entry is approximated using the average of entry values from all other monitoring stations that are not missing.

Visualize Pollutant Values for Each Monitoring Station

To visualize each monitoring station's data, a horizontal bar chart was implemented that can help users to better perceive the underlying meaning of the data. Users are able to choose which monitoring station to visualize, and which date to visualize by using the controls provided in the main displaying area, as shown below. After the user picks a date or a station, the title that indicates the daily average AQI and the bar charts of 6 types of pollutants would be updated accordingly. The monitoring stations are now provided with their station codes. These codes will be matched against the location data file and converted into English names in the next stage.



The daily average values of 6 types of pollutants have been already provided in the data files, which are the 24h average at 23:00 on each day. The daily average of AQI is not originally provided in the data files. Its value was obtained by summing up the value of each hour, and divided by 24 hours.

Because 6 pollutants do not have the same unit, and they are weighted differently in calculation of AQI, it does not make sense to provide axes on the graph. Instead, each bar was drawn more like a progress bar. Each bar was drawn by using linear scaling, meaning it has the domain from minimal value to maximal value of each type of pollutant, and has the range from 0 to the maximal pixel value of the bar with on the screen. By scaling the values in this way, the midpoint of each bar indicates the position for the daily average value of that type. Thus, self-

comparison can be easily made. For example, as the image shown in the Interface section, the value of PM_{2.5} exceeds the midpoint of the bar, which indicates that this location has a high PM_{2.5} pollution comparing to the average value across the nation. Similarly, SO₂ pollution does not “progress” too much, which means SO₂ is not the main pollution around this location.

Current Findings

Different locations do have different pollution patters. For instance, the monitoring station 1001A has extremely high PH_{2.5} pollutant comparing to other pollutant, and the monitoring station 2566A has extremely high O₃ pollution, which implies that the assumption about different pollution pattern may exist. The further examination on pollution patterns will be made through k-means clustering and choropleth map.

In addition, a few monitoring stations’ values were not correctly recoded. For example, by examining the maximal values, it was found that the monitoring station 1742A has an astonishingly high O₃ 24h-average value, 1200 µg/m³, while other values of this type are all approximately ranging from 100 to 200. It is impossible to have such a high value. To keep the accuracy of the data, these incorrect entries were all replaced with the average value of its type.

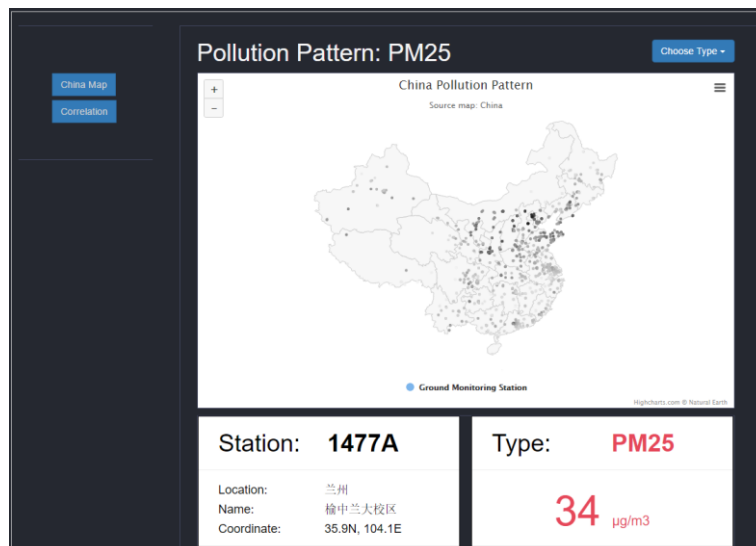
	YJ	YK	YL	YM	YN	YO	YP	YQ	YR	YS
349	111	104	107	119	62	62	65	58	47	47
350	216	247	314	230	162	147	150	202	124	108
351	151	158	205	168	91	77	91	97	102	70
352	102	117	99	100	27	10	30	60	32	30
353	78	76	69	80	28	13	36	33	26	13
354	54	56	77	44	100	74	96	87		8
355	39	35	41	43	52	35	37	40	12	15
356	44	40	13	67	7	10	4	11	9	148
357	99	122	103	135	147	175	165	124	1200	180
358	66	81	60	90	79	90	101	68	199	150
359	88	110	93	125	134	156	152	110	298	178
360	3.2	3	2.9	3.2	1.7	1	1.7	2.9	1.2	0.9
361	2.1	1.9	2.3	1.9	1.3	0.8	1.1	1.8	1	0.7

Final Report

China Map Interface

Besides the bar chart aforesaid, the project also provides an interactive China map. The map is zoomable, so that users can zoom in and focus on a certain region that they are interested in. Each ground monitoring station is represented by a small circle, and correctly placed in the map based on its real-world longitude and latitude. The color of each monitoring station is linearly interpolated between the maximal and minimal value of the current pollutant type. Thus, the darker colored stations have relatively higher pollutant values, and the lighter colored stations have lower pollutant values.

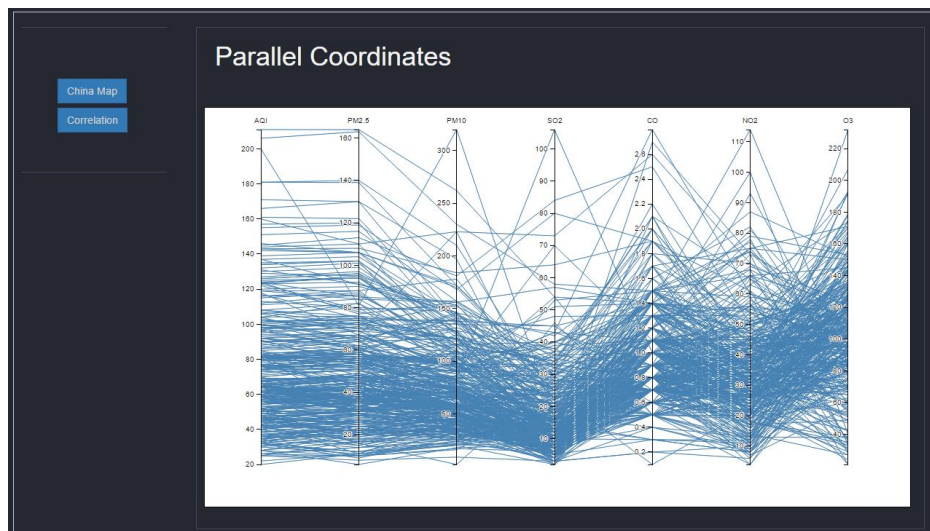
The station circles in the map are clickable. Once users click on a station circle, the panel on the bottom will change accordingly to show the details on that station. The bar chart on the bottom will also change accordingly. The bottom left panel displays the station code, real-world coordinates, the city where the station is located (in Chinese), and the station's name (in Chinese). The panel on the bottom right indicates the current pollutant type and the daily average of such pollutant. The drop-down list on the top right can change the pollutant type to examine.



The picture above demonstrates the PM2.5 pollution pattern over all 1487 monitoring stations, and the panel on the bottom shows the detail of Station 1477A.

Parallel Coordinate plot

The parallel coordinate plot provides an interactive way to visualize and analyze the correlations among Air Quality Index and six types of pollutants. The correlation between each pair of axes are calculated in advance to determine the arrangement of axes in the plot. The most correlated axis pairs are placed to be adjacent to each other. Since there are too many data to be concerned, data reduction is performed before plotting. It is assumed that the adjacent monitoring stations in the data file are also very close to each other in the real world. Their recorded values do not vary too much due to proximity. Therefore, for every three monitoring stations, only one of them is picked and drawn in the plot. The following pictures shows the parallel coordinates after data reduction.



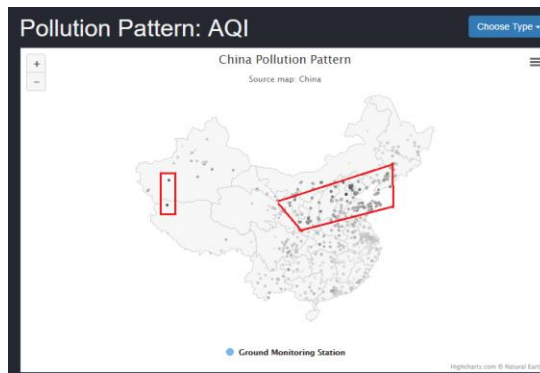
The parallel coordinate provides brushing functionality on each axis, so that users can focus on a portion of the total data.

Findings

1. Pollution Patterns

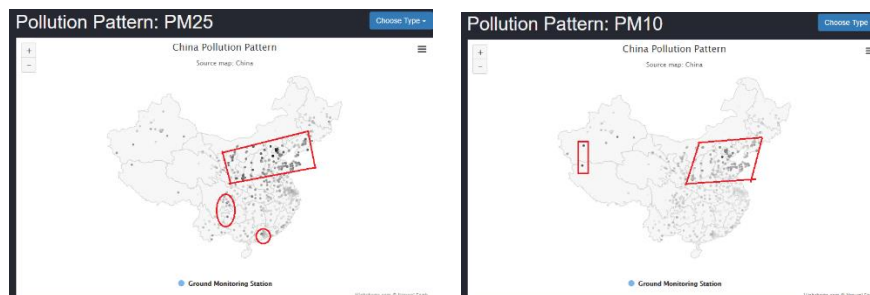
1.1 AQI Pattern

Air Quality Index indicates the overall air quality of a certain region. The map shows that the most polluted areas are in the north, roughly covering Beijing, Tianjing, Shandong province, Shanxi province, Hebei province, and Liaoning province. The air quality of western Xinjiang Province is also lower-ranked.



1.2 PM2.5 and PM10 Pollution Patterns.

PM2.5 and PM10 are solid particle matters that adversely influence the air. These two types show the closest correlation with AQI. Their pollution patterns mostly agree with that of AQI.



For PM2.5, there are two more pollution centers: Southeastern Sichuan Province, which is a region producing ores and fossil fuels, and Zhu River Delta region of southern Guangdong province.

1.3 SO2 Pollution Pattern.

SO2 is another pollutant that is closely correlated to AQI. It has similar pollution pattern of that of PM2.5.



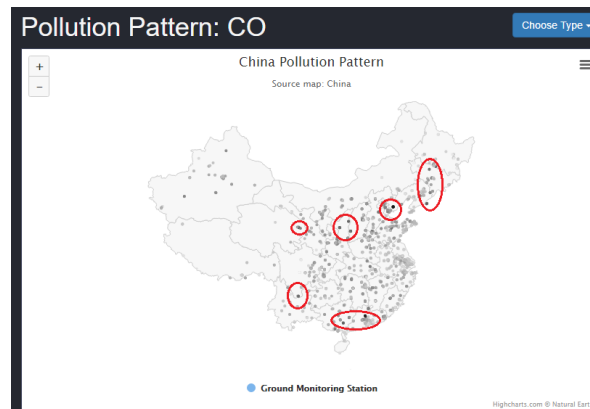
1.4 NO2 Pollution Pattern

NO2 pollution pattern is slightly different from those above. The most NO2 polluted areas are along the east coast and in Sichuan province where the population is dense. Thus, the pattern here may imply NO2 pollution is related to human activities.



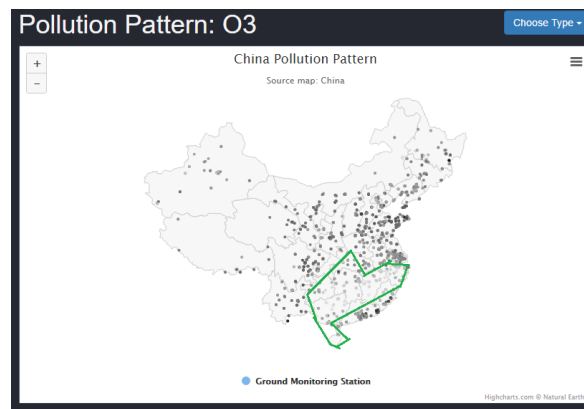
1.5 CO Pollution Pattern

The CO pollution centers are scattered all over China. These centers are located near industrial cities, showing a strong relationship between CO pollution and industrial production.



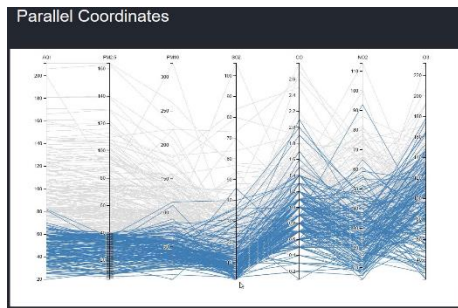
1.6 O3 Pollution Pattern

O3 pollution shows a quite different pattern from others. It dominates all the regions except the region circled out by the green lines.

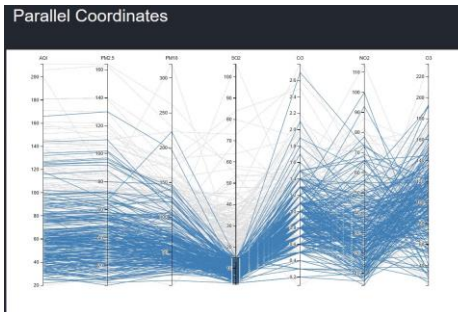


2. Pollutant Correlation

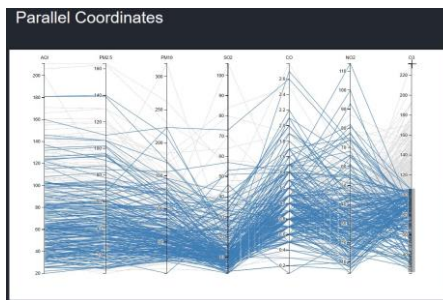
2.1 AQI is mostly correlated with PM2.5, PM10, and SO2. AQI is directly proportional to PM2.5, PM10 and SO2.



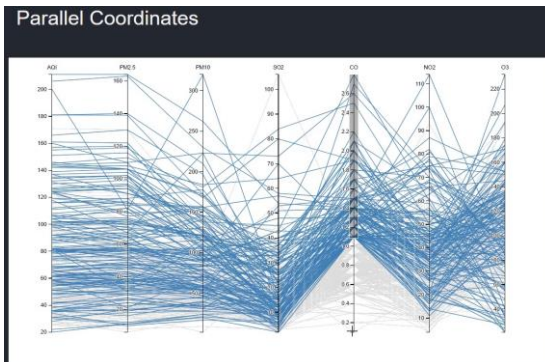
2.2 Low SO2 pollution value is not enough to imply a better air quality.



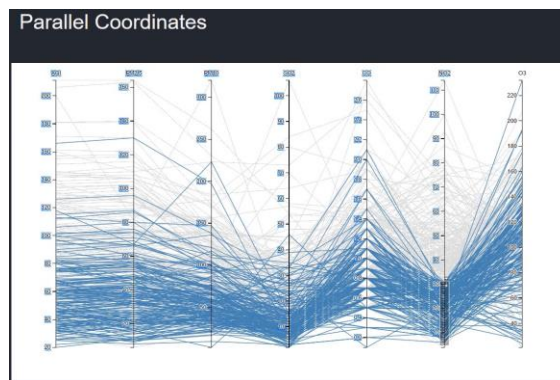
2.3 O3 value is not clearly correlated to any other pollutants.



2.4 High CO pollution indicates high NO₂ pollution. However, CO solely does not necessarily determine AQI.



2.5 Most of good AQIs do have lower NO₂ values. Also, low NO₂ values imply low SO₂ pollution.



Conclusion

The bar chart in this project provides an easy way to make cross-comparisons for each pollutant, and spot which pollutants are dominant in the proximity of a certain monitoring station. The interactive China map demonstrates pollution patterns for each type of pollutant. The Parallel Coordinate Plot makes it easy to examine and analyze the correlations among AQI and six types of pollutants.

The most polluted regions are in the north of China, surrounding the Bohai Sea. The air quality is mostly determined by PM2.5, PM10, and SO2, among which PM2.5 is the closest correlated. The pollutant NO2 seems to be more related to human activities since its pollution centers are in the regions that have high population density, and NO2 shows a direct proportion to SO2. CO pollution centers are scattered all over China, showing a strong relationship with industrial production. O3 pollution is independent, which has no correlation with any other pollutants, and its pollution pattern hardly agrees with those of others.

References

1. Plain China Map API, *highcharts.com*, <https://jsfiddle.net/gh/get/library/pure/highslide-software/highcharts.com/tree/master/samples/mapdata/countries/cn/cn-all>. Accessed on May 12, 2017.