# AQI analysis for the China Mainland between 2000-2015

*Tian Xia*

*10/12/2019*

## 1. Introduction

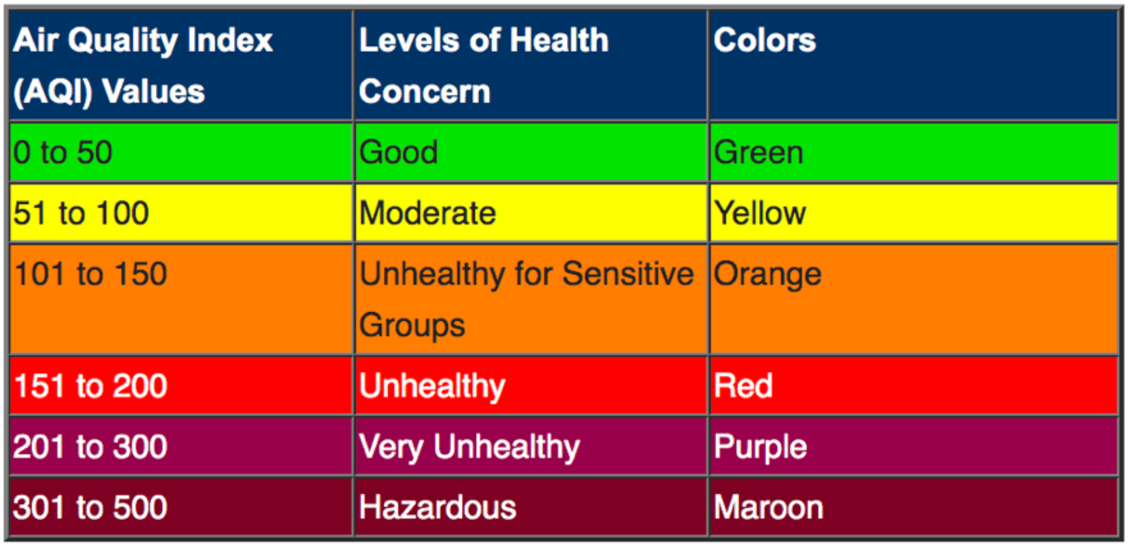
Air pollution occurs when harmful or excessive quantities of substances, including gases, particles, and biological molecules are introduced into the Earth's atmosphere. It may cause diseases, allergies, and even death to humans; it may also cause harm to other living organisms such as animals and food crops and may damage the natural or built environment. Both human activity and natural processes can generate air pollution. We always use the API(air pollution index) to describe the level of air pollution's degree. The AQI is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. In some countries, air pollution gradually became a serious problem. In Beijing, the capital of China, the average API of the whole year, is 150, and it is in the unhealthy range of the API standard.

In this project, my goals are to create the prediction model for the AQI data in each city of China. Based on my models, I can know the influence of each factor to the AQI for each city and to predict the tendency for each cities' AQI level variance. Finally, I hope my model can help the government and citizens to improve the air quality in high AQI level cities. In my project, there was no generally clients. However, the API is a severe environmental problem in China; most people have interested in it includes citizens and government. For the government, this project may help them to decide how to decrease air pollution. For citizens, I hope they can know how dangerous the API is, how it affects human health, and how to avoid it. And also, I hope my project can advocate for most people in China to do something to decrease the API.

age

## 2. Background

The data source is from Ming Cheng's project.[1], which is a public project created by author Ming Cheng. The project includes every China mainland cities' AQI data for 2000-2015. The author scrape the data from the Ministry of Ecology and Environment of the People's Republic of China.[2] The AQI data set include five kinds of pollutions in the air: Tropospheric Ozone(O3), Particulates, Sulfur Dioxide(SO2), Carbon Monoxide(CO), Nitrogen Dioxide(NO2). However, in this project, we don't consider each element of air quality separately. The Ministry of Ecology and Environment of the People's Republic of China (MEG) already detected and collected all of them to calculate the AQI in every city of China mainland. The government calculates AQI data hourly, then SO2, NO2, CO concentrations are measured as average per 24h, O3 concentration is measured as an average per hour, and the moving average per 8h, PM2.5, and PM10 concentrations are measured as an average per hour and per 24h. This more details for how AQI calculated are on the airnowtech site.[3]



| Air Quality Index (AQI) Values | Levels of Health Concern | Colors |
|---|---|---|
| 0 to 50 | Good | Green |
| 51 to 100 | Moderate | Yellow |
| 101 to 150 | Unhealthy for Sensitive Groups | Orange |
| 151 to 200 | Unhealthy | Red |
| 201 to 300 | Very Unhealthy | Purple |
| 301 to 500 | Hazardous | Maroon |

age

The AQI divide by 6 levels from low to high air quality. We can see the difference in the following photo. Unfortunately, In China, most cities' AQI is in yellow and orange. Therefore, the AQI problems are urgent to solve for China government.

# 3. Method

The data set I use is basic from Ming Cheng's data set. He scrapes the data from the government website. However, the government doesn't do the categorization by time and cities' names in the raw dataset. Therefore, the raw dataset has 444391 data, but the simplified version of the dataset is created by data processing for future data analysis. I choose RStudio as my tool to do the data analysis because I have already studied how to use it for one year. I have experiences in data analysis by RStudio. As we all know above, there have many factors that can affect the air quality, include factory, automobile exhaust. We don't want to create a model for all of them. Therefore, finding the target factors is essential for my prediction model. Firstly, in this semester, I use R to do the cluster analysis for my dataset, including K-mean Cluster and Cluster Dendrogram. Through doing the cluster analysis for my dataset, I can classify the data to several groups based on the pollution level, then find the possible protentional relationship(factors) between the cluster. For example, high AQI cities are probably near the heavy industrial area so that the heavy industry may be an essential factor in the air quantity. In the first semester, my goals are to find the target factors by doing the cluster analysis. After this semester, I plan to use these factors to do the concrete analyze for the data. The linear regression model is created for these factors by R. To do this, and I will research more data include each factors' variation between 2000 and 2015. Then, create a weighted algorithm for each factor by linear regression, to find which factors are the most critical factors to the AQI.

# 4. Timeline and some protentional problems

# Fall October:

I plan to finish all data processing and start to do the cluster analysis. On recently two weeks, I focus on to create a simplified version of the raw dataset. The raw dataset has 444391 data, and it is a huge dataset and hard to do the data analysis on it. The data processing is the first thing I need to do. As I mentioned above, the raw dataset doesn't classify data by cities and time. The first thing I did is to calculate the average API for the whole year between 2000-2015. So, I can analyze the data on the year. After that, I classify the raw dataset based on the city name and year. Finally, I get a new 32*20 dataset. It is what I will use in the following data analyze. Even I already used R to do the data analysis several times, but I didn't have much experience in cluster analysis. Therefore, in October, studying is also an important process on my timeline. I already found some libraries and videos about cluster analysis. I will study them all and try to start use them. And also, keep bringing the question I meted to discuss with classmates and professors.

# Fall November:

I hope I have learned all the necessary things about cluster analysis in October and start to analyze the dataset. After November, my goal is to finish the cluster analysis and found the target factors, and also including a little data visualization works. And Try to start to write draft of the proposal. I will write my final report by R markdown, including the basic formula and figures.

# Fall Decembers:

At this time, I need to finish all my works and submitted the final report.

## Spring TimeLine:

Researching the new dataset for my target factors is important. It will be a huge work. I probably spend the most time researching data and data processing in my first 1-2 months. After that, the regression model will be created to predict the influence of the factors on AQI data. And finally, prepare my final report and presentation for this senior capstone project.

## 5. References

1.Ming Cheng's project: https://github.com/mingcheng/AQI

2.Ministry of Ecology and Environment of the People's Republic of China: http://datacenter.mep.gov.cn/

3.airnowtech site:https://forum.airnowtech.org/t/the-aqi-equation/169.