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 caner 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 sever adverse health effects. The AQI is mainly depend on 6 types of airborne pollutants, PM2.5, PM10, SO_2, NO_2, O_3, and CO. Therefore, it is important to find how each type of pollutant is correlated with AQI recently. The objectives are following:

1. Find the nation-wide correlation between AQI and each type for March 2017 and visualize correlations with scatter plot matrix.
2. Visualize hourly value of 6 types of pollutant and AQI with parallel coordinate plot, and search for the correlations.
3. All monitoring stations will be grouped by their city locations. For each city, visualize monthly average of 6 types of pollutant for all monitoring stations in that city in bin chart.

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

4. Find the pollution centers and clusters for each type of airborne pollutant with k-means clustering algorithm.
5. Visualize the geographic distribution of each type of air pollutant with choropleth map.

Timeline

The project will proceed in a bi-weekly manner.

Stage 1	April 12 ~ April 25	Complete Objective 1 to 3. Prepare for preliminary report.
Stage 2	April 26 ~ May 12	Complete Objective 4 and 5. Overall testing and integration. Prepare for the final report and presentation.