U.S. Pollution Dataset Analysis

Peng Yan
Crew
University of Colorado Boulder
peya8291@colorado.edu

Zhixing Chang
Crew
University of Colorado Boulder
chzh5137@colorado.edu

Tianlun Zhao
Crew
University of Colorado Boulder
tizh6070@colorado.edu

Xiaoyang Dou
Crew
University of Colorado Boulder
xido3947@colorado.edu

ABSTRACT

Air pollution is one of the most important environmental issues. Economic development, urbanization, energy consumption, transportation/motorization, and rapid population growth are major driving forces of air pollution in large cities, especially in megacities. Air pollution levels in developed countries have been decreasing dramatically in recent decades. However, in developing countries, air pollution levels are still at relatively high levels, though the levels have been gradually decreasing or have remained stable during rapid economic development. The World Health Organization (WHO) published the "WHO Air Quality Guidelines (AQGs), Global Update" in 2006. These updated AQGs provide much stricter guidelines for PM, NO2, SO2 and O3. Considering that current air pollution levels are much higher than the WHO-recommended AQGs, interim targets for these four air pollutants are also recommended for member states, especially for developing countries in setting their country-specific air quality standards [1]. This dataset deals with pollution in the U.S. Pollution in the U.S. has been well documented by the U.S. EPA. We try to use this data find answers about some question. For example, what is the main source that cause air pollution in different city? How does the trend of air pollution during the day? What are the main pollutants in each city? which cities' pollution is getting better? which cities' is getting worse?

1 INTRODUCTION

With rapid economic development, people start to pay more attention to health issues. Bad air condition can result in uncomfortable feelings in eyes, nose, throat and lung area. It can not only cause disease like aggravate asthma and other respiratory conditions, it might even damage heart and cardiovascular system. Since everybody breathe, it is important to monitor air condition so people can protect themselves when necessary.

Four important chemical matters that are the source of the air pollution are NO2, O3, SO3 and CO. NO2 is Nitrogen dioxide, a nasty-smelling gas. It is naturally formed by lighting, and some are produced by plants, soil, and water. However, not all nitrogen dioxide in our country's air are formed in this way. The major part of the nitrogen dioxide in our atmosphere is generated by burning fossil fuels like coal, oil, and gas. Since our cars run on fuels, motor vehicle emission is one of the major causes of excess nitrogen dioxide gas. Some other source is manufacturing industries' everyday production activities, such as food processing, electricity and heat generation from coal-fired stations. This gas irritates respiratory tube. Long time exposure to it increases the risk of lung infections, respiratory conditions, higher response to allergens and even heart disease.

O3 is ozone, a gas compound with three atoms of oxygen. It exists in both upper atmosphere and ground level. Ozone at ground level is a harmful air pollutant. It is blamed for the ozone pollution. This kind of air is like nitrogen dioxide, that generated by cars, power plants, industrial activities and some other chemical reactions that happen in sunlight. Breathing ozone can also trigger various health problems like nitrogen dioxide does, with symptoms like coughing, scratchy throat and chest pain.

SO3 and CO are like the two we talked above, also threat people's health. Therefore, we are interested in monitoring these gas indexes in our air. The dataset we choose to process is about these four pollutants' everyday condition in U.S. from 2000 to 2016. Each pollutant's max value and max hour is documented. Other attribute types in our dataset include monitoring site(address, city, state, country), and monitoring date code. By relating the max value with monitoring site, we can see the trend and pattern of

these pollutants condition in some particular address, city and state. We can also relate max hour with monitoring date or relate one pollutant value to another one. There are many possibilities with this dataset. Depend on the goal we have, we can make visualizations to help achieve it. And furthermore, by visualizing information, we can think deeper like what the reason of the increase or decrease of certain pollutant is, what the cause is, and what we learn from this experience.

2 MOTIVATION

It is common for developing countries to suffer from air pollution. London used to be known for its London fog. California also deals with air quality issue for many years. Our data includes 28 fields (data type include numeric and nominal/string), four major pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016. U.S. as the developed country went through the bad air condition phase, and it has experience dealing with air pollution. By analyzing the U.S. past 16 years air condition data, we can learn from it. By digging into the data, we can see which place has changed from old air condition, and by doing more research, we want to know what kind of methods government used or regulations help better air quality, or what kind of activities worsen the circumstance. In addition, we can conclude the treating methods for other countries that is or will be the victim of the air pollution.

3 PREVIOUS WORK

There are many data research about air pollutants in recent years. For example, “openair — an R package for air quality data analysis”. Since 2011, frequent occurrences of haze in China have become a cause for panic and routinely appear as a major topic in the media and on climate websites. Visual exploration of air pollution with spatio-temporal data is a solution that makes complex data understandable because graphical representation is relatively intuitive. However, there are several problems that prevent the widespread use of more insightful analysis. For example, a coherent set of data analysis tools for air pollution purposes does not exist. and, many users are not aware of the tools available or how to apply them. If these problems can be overcome, there are many potential benefits including: a more comprehensive evidence base to support decision making, identification of the factors controlling pollutant concentrations [2]. Through another research paper “Air pollutants measured in Seoul”, we can find the researcher did a lot of data clean work, filling missing values with 0s, filling missing values with mean, filling missing values with interpolation. We also can visualize the data set that could help us do some analysis work [3].

Compare the previous research our data analysis focus on how to help user in different place of U.S. quickly and accurately to find what are the major pollutants and what are the relationship between of them. Then, it can guide people or government formulate corresponding policies.

We have done few tasks. So far, we already derive some different pollutants’ average value graphs. These graphs could help user visualize information. Users also can easily find the information that they need. For example, if users want to know A certain pollutant trend in recent years, we can through the line graphs see the trend. We can see it from figure 1 and figure 2.

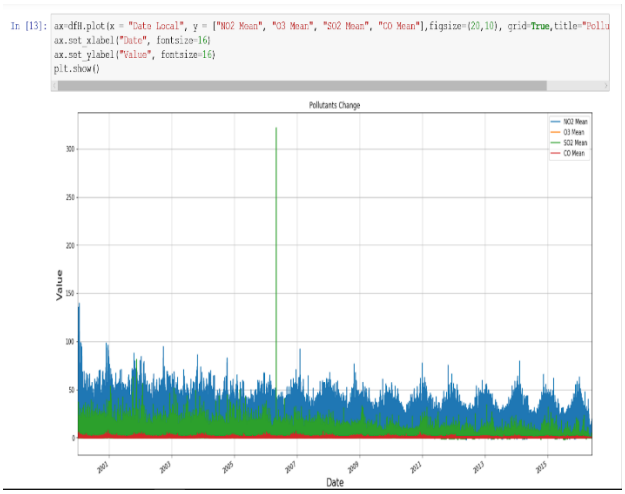

Figure 1

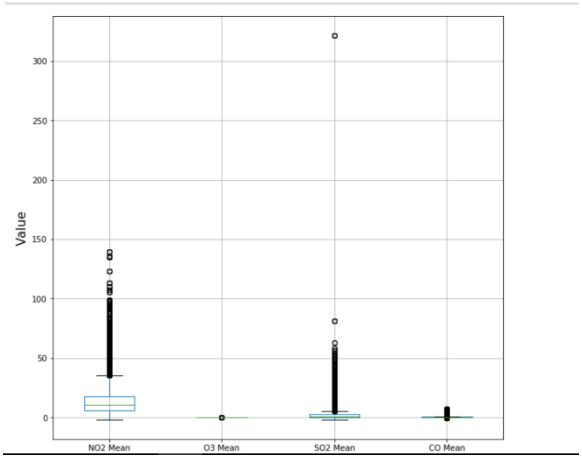

Figure 2

4 PROPOSED WORK

Data collection
We found the U.S. Pollution data from Kaggle website by Brenda. There are more than millions of data points in this data set. This is the main source we will use. Then, we are going to find some useful materials about pollutants to prove our analysis.
Data Cleaning
The data sets from U.S. Pollution data filtering algorithm and stored in an Excel format. We need to extract these data to the form that we will use. This step will reduce after manipulation of the attribute subset of the data set.
Data Reduction
The U.S. Pollution data of the original data set according to time order to show the value of pollution change. Every object has a lot of different properties, some of them is unnecessary for this project. The first step in the data reduction is to reduce the number of attributes. Such as, state code and country code. We clearly know the dataset about the U.S. and the specific name of the state and city, so we don't really need the state code and county code. We can integrate them.
Transforming
We need to use these data set to make pollution tendency graph that can help us to analysis the pollution problem in The United States and it looks more intuitive and easier to understand.
Histogram
The stable factors are the arithmetic mean of concentration of pollutants and the calculated air quality index of pollutants from 2000-2016 in different cities of U.S.
Evaluation
Based on the results of our study we can determine the relationship between different pollutants. We through do some trend chart and see what has happened to increase or decrease pollution and some cities make pollution to increase or decrease. Using this result we can guide different cities use different methods to reduce pollution.

5 DATA SET

Our data is coming from The U.S. Pollution Data. It includes State Code, County Code, Site Number, Address, State, County, City, Data Local, NO2 Units, NO2 1st Max, NO2 AQI(The calculated air quality index of NO2 within a given day), O3 Units, O3 Mean, O3 1st Max, O3 AQI(The calculated air quality index of O3 within a given day), SO2 Units, SO2 Mean, SO2 1st Max, SO2 AQI(The calculated air quality index of SO2 within a given day), CO Units, CO Mean, CO 1st Max, and CO AQI(The calculated air quality index of CO within a given day).

Through this dataset, we can have detailed understanding every American city of main pollutants and specific pollution index. Based on the analysis of these data we can get the pollution of heavy trend of a city from 2000-2016. Through our team's textual data analysis and chart trends, decide whether this city should focus on pollution sources and take appropriate measures.

6 TOOLS

The whole data is in csv file, so the first tool will be used is excel which contains tables and many programs that can help us get some main ideas about the whole data. Excel is frequently used, so it will take much less time than other kind of tools to help us getting ideas about this "US Pollution" dataset. However, because of the huge amount of the dataset, only using excel is not enough for analyzing. Python 3 and Weka are needed for analyzing the dataset more detailed. We also need to analyze the whole dataset from different aspects. We are going to use Python3 to with the formula to get the z-score, correlation and so on. Tableau is a tool for making visualizations. Tableau can make many different kinds of graphs which can be really creative so that readers can view and easily understand the content of the graph and the data set. However, the amount of the dataset is more than a million, so we are not sure if the Tableau can work well as it dealing with the small amount number of data. Therefore, we may consider about using python 3 to make graph as a back plan.

7 MILESTONES

After this part of the project, we will finish researching of the past report about the "US Pollution" dataset, and we will get the main idea about the part and ways we want to analyze this dataset.
·Data clean: we want to clean the data first.
·Data reduction: after clean the data, we will separate the data into different parts and assign each one a part of data
·Data transformation: each of us will deal with their part and get the result of data on the relationship between different attributes and the correlations.
·Deeper data mining: after knowing the correlations, we will try to deal with the attributes which are correlated.
·Analyzation and Evaluation: evaluate the result we got.

8 SUMMARY OF PEER REVIEW SESSION

As international students, grammar is a big problem for us. When we write the report, we need focus on grammar. For the data integration part, we need to focus on detecting and resolving data value conflicts. For the data processing we need to focus on solve specific problems.

9 REFERENCES

1. Environ Health Prev Med. 2008 Mar; 13(2): 94–101. Air pollution and population health: a global challenge
Published online 2008 Feb 28. doi: 10.1007/s12199-007-0018-5
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698272/
2. openair — an R package for air quality data analysis David C. Carslaw,∗, Karl Ropkins
King's College London, Environmental Research Group, Franklin Wilkins Building, 150 Stamford Street, London SE1 9NH, UK
Institute for Transport Studies, University of Leeds, LS2 9JT, U
http://www.openair-project.org/pdf/openair_paper_preprint.pdf
3. Air pollutants measured in Seoul
https://www.kaggle.com/jihyeseo/seoulairreport