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

Zening Li
Crew
University of Colorado Boulder
zeli6634@colorado.edu

Yifan Li
Crew
University of Colorado Boulder
yili4303@colorado.edu

ABSTRACT

Air pollution is one of the most critical environmental issues. Economic development, urbanization, energy consumption, transportation/motorization, and rapid population growth are major driving forces of air pollution in large cities. Air pollution levels in developed countries have been decreasing dramatically in recent years. 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, Nitrogen Dioxide, Sulfur Dioxide and Ozone. 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 are trying to resolve several questions about air pollution by using this dataset. For example, what is the main source that causes air pollution in the different city? How does the trend of air pollution during the day? What are the main pollutants in each city? How is the air pollution problem in different cities? Are those cities having better air quality, or worse? We create some graph focus on four major pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) change trend. Through our graph, we can see the AQI Mean in All U.S. States Each Year From 2000 to 2016 became better year by year. Four major pollutants 1st max value over time (from 2000 to 2016) also decrease in recent years.

1. INTRODUCTION

With rapid economic development, people start to pay more attention to health issues. Lousy air condition can result in uncomfortable feelings in areas of eyes, nose, throat, and lung. It can not only cause disease like aggravate asthma and other respiratory conditions, but it can also damage the heart and cardiovascular system. Everyone breathe everyday, therefore it is essential for people to monitor air quality so that people can protect themselves when necessary.

The air pollution contains four major kinds of gases, which are Nitrogen Dioxide, Ozone, Sulfur Dioxide and Carbon Monoxide. Nitrogen dioxide is 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 is 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 of exposing to it can increase the risks of having lung infections, which can cause the problem of respiratory, higher response to allergens and even heart disease which can be the main problem to death.

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

Sulfur Dioxide and Carbon Monoxide are like the two above we talked about, because of human activities, this two kind of gases are generated and emitted over the value which it should be. They also threaten people's health.

Due to how harmful these air pollutants are to human beings, 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 the U.S. from 2000 to 2016. Each pollutant's max value and max hour is documented. Other attribute types in our dataset include monitoring all state and monitoring date code. By relating the max value with monitoring site, we can see the trend and pattern of these pollutants condition in all state. We can also relate max hour with monitoring date to each pollutant or relate one pollutant value to another one. Within this dataset, there can be many possibilities we can discover and evaluate. Depending on the goal we have which is trying to have some kinds of solutions for the pollution, 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].

2. RELATED 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. We think we could use D3 build a graph, that also can overcome the problem[3].

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[4].

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.

The second paper comes from Dr Gerard Hoek, Bert Brunekreef, Sandra Goldbohm, Paul Fischer, and Piet A van den Brandt. 19 October 2002. Association between mortality and indicators of traffic-related air pollution in the Netherlands. The background of this paper is "Long-term exposure to particulate matter air pollution has been associated with increased cardiopulmonary mortality in the USA. We aimed to assess the relation between traffic-related air pollution and mortality in participants of the Netherlands Cohort study on Diet and Cancer (NLCS), an ongoing study"[5]. This paper is about a experiment which "investigated a random sample of 5000 people from the full cohort of the NLCS study (age 55–69 years) from 1986 to 1994. Long-term exposure to traffic-related air pollutants (black smoke and nitrogen dioxide) was estimated for the 1986 home address. Exposure was characterised with the measured regional and urban background concentration and an indicator variable for living near major roads. The association between exposure to air pollution and (cause specific) mortality was assessed with Cox's proportional hazards models, with adjustment for potential confounders"[6].

This paper is useful for our project. The reason is that we wish to discover the data of air pollutants in the USA. If we only look at the data, there is only numbers in it, so we need some experiment as a background. This paper is a really good experiment background for our project. Because it also aims air pollution. Also, this paper shows the relationship between air pollution and traffic, which I think it is important to our data. The reason is that we want to shows some results, which some of the air pollutants come from traffics.

The second paper comes from Shiri Avnery, Denise L. Mauzerall, Junfeng Liu, and Larry W. Horowitz. April 2011. Global crop yield reductions due to surface ozone exposure: 2. Year 2040 potential crop production losses and economic damage under two scenarios of O3 pollution. This paper is mainly about "We examine the potential global risk of increasing surface ozone (O3) exposure to three key staple crops (soybean, maize, and wheat) in the near future (year 2030) according to two trajectories of O3 pollution: the Intergovernmental Panel on Climate Change Special Report on Emissions Scenarios (IPCC SRES) A2 and B1 storylines, which represent upper- and lower-boundary projections, respectively, of most O3 precursor emissions in 2030. We use simulated hourly O3 concentrations from the Model for Ozone and Related Chemical Tracers version 2.4 (MOZART-2), satellite-derived datasets of agricultural production, and field-based concentration:response relationships to calculate crop yield reductions resulting from O3 exposure. We then calculate the associated crop production losses and their economic value. We compare our results to the estimated impact of O3 on global agriculture in the year 2000, which we assessed in our companion paper [Avnery et al., 2011]. In the A2 scenario we find global year 2030 yield loss of wheat due to O3 exposure ranges from 5.4 to 26% (a further reduction in yield of +1.5–10% from year 2000 values), 15–19% for soybean (reduction of +0.9–11%), and 4.4–8.7% for maize (reduction of +2.1–3.2%) depending on the metric used, with total global agricultural losses worth $17–35 billion USD2000 annually (an increase of +$6–17 billion in losses from 2000). Under the B1 scenario, we project less severe but still substantial reductions in yields in 2030: 4.0–17% for wheat (a further decrease in yield of +0.1–1.8% from 2000), 9.5–15% for soybean (decrease of +0.7–1.0%), and 2.5–6.0% for maize (decrease of + 0.3–0.5%), with total losses worth $12–21 billion annually (an increase of +$1–3 billion in losses from 2000). Because our analysis uses crop data from the year 2000, which likely underestimates agricultural production in 2030 due to the need to feed a population increasing from approximately 6 to 8 billion people between 2000 and 2030, our calculations of crop production and economic losses are highly conservative. Our results suggest that O3 pollution poses a growing threat to global food security even under an optimistic scenario of future ozone precursor emissions. Further efforts to reduce surface O3 concentrations thus provide an excellent opportunity to increase global grain yields without the environmental degradation associated with additional fertilizer application or land cultivation."[7].

This paper is useful to our project because from the O3 visualization, we see the trend is flat compared with other pollutants, so we think O3 is a special air pollutant compared with other air pollutants. We wish to discover more about O3 pollutant in the future. This article shows the effect of O3 to crops, which I think it is a good resource to our future discover. Also, since the article says it is not a risk to current world, so we think this is a reason why people does not control O3 very immediately.

In a paper called “ambient air pollution WHO report”, the article points out that there are approximately 3 million people died from ambient air pollution. Global map of annual mean concentration of PM2.5 μg/m3. This shows that the regions of the world that hold most of the world’s populations are those that are exposed to PM2.5 levels that are above the annual mean WHO quality guideline levels. 94% of deaths worldwide are due to noncommunicable diseases in adults, such as cardiovascular diseases (stroke and ischaemic heart disease), chronic obstructive pulmonary disease and lung cancers. Additionally, one in ten people lives in a city that complies with WHO Air quality guidelines. Furthermore, two main culprits of PM2.5 are black carbon (soot) and soil dust[8].

After I read this paper, I was very upset that there exits that many people are still suffering from no cleaning air. Comparing to the analysis we have done, we can clearly see a huge difference between the regions where this paper mentioned and United States. The aim for WHO is helping people who need help. But, the effort that WHO can do is limited. Thus, I believe that what WHO wants is to appeal people help people from developing countries together.

There is another paper I fond called “atmospheric black carbon in PM2.5 in Indonesian cities” . The main idea is about a case study done in Manila for impact on vehicle emission. Manila is one in the top twenty most polluted cities in the world and PM2.5 is 70% higher than WHO recommended level. Additionally, 1 in 4 deaths is caused by ambient air pollution in Manila

Increases respiratory issues and is linked to cancer/stroke. Among all pollutions, 80% of air pollution is caused by vehicles in Manila. For solving this problem, Indonesian government has three plans to do. Firstly, better monitor air quality. Second, to require more stringent automobile emissions testing and increase inspections. Last but not least, encouraging more public transit and cleaner vehicles[9].

Vehicle emission is kind of the problem that is very hard to avoid in most of developing countries. In order to develop, most developing countries are not able to afford cost that take air pollution into a consideration. Unlike some developed countries, United States for example, they have extra budget for protecting environment. Thus, we can drive that most countries that has air pollution problem are from developing countries. I believe that developed countries should spend more money and budget on helping developing countries with pollution problem.

We found a paper called “human settlements and environment commission of Shenzhen municipality”. The main idea is that China is working on cleaning air. Many regulations and restrictions have been published in Shenzhen since 1990s. For example, Routine testing and inspection of pollution sources, “Blue Sky Action” and Power plants, boilers, vehicles and dust control[10].

Shenzhen is the third largest city in China and Shenzhen doesn’t has a serious air pollution. This paper mainly focuses on why Shenzhen could achieve that. This is because that most of companies in Shenzhen is technology orientated. And Shenzhen government also spend a lot of money on plants and park. Comparing to the United States, the only problem is that Shenzhen has way more population than any city in the United States. This will also cause a certain portion of pollution by vehicle emission. But, comparing to some cities that has heavy industry, Shenzhen has more advantages on air pollution problem.

We found a paper called “Taiwan: Action Plan to Reduce Air pollution, Shift to Electric”. The main idea is that, 66% is from domestic sources, 34% is from the Asian dust storms that blow in from China the spring. The government of Taiwan has passed the Air Pollution Control Act and created an action plan to shift to electric. Government will need to do more. Motor vehicles are
only 30% of the problem. The government will have to focus on establishing policies for fix emitters and the construction sites[11].

Electrical vehicle can definitely reduce a lot of air pollution. But, the most serious problem is that a lot of countries cannot support large amount of super charger for these electrical vehicle. Some of these countries cannot even afford electrical vehicle. Thus, this is not possible for some developing countries to reduce air pollution using electrical vehicle. In the United States, electrical vehicle starts boom in recent years.

3. Description and Justification

we began with cleaning and preprocessing the data to get the million data entries into a more manageable and usable form.

The initial U.S. Pollution Data was fairly messy and had a lot of missing data and other various issues. There was over 1.4 million dataset that gathered four major pollutants (Nitrogen Dioxide, Sulfur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016.

In the U.S. Pollution data, there is a total of 28 fields. The four pollutants (Nitrogen Dioxide, Ozone, Sulfur Dioxide and Ozone) each has five specific columns. It needed a lot of work. Through use the box plot (Figure 1), we found there were many missing values and outliers recorded by the million data of four major pollutants (Nitrogen Dioxide, Sulfur Dioxide, Carbon Monoxide and Ozone). Most of the data were probably automatically recorded daily by data instruments which maybe make mistakes occasionally. Some of the missing values and errors were easily removed/ignored because the data usually
included has significant outliers. However, there were
many datasets that appeared normal at a glance, but actually, they were outliers that could severely alter our analysis. In some instances, the data had values like N/A indicating no data. In other instances, which were much more difficult to detect, was when a data value for something like the maximum daily Sulfur Dioxide mean value equaled something like above 300 (Figure 2). Comparing with the other data, this data is a bad value and should be ignored, but when having to work with all the various data types, it became increasingly difficult to have to detect these noisy data values. It was not as simple as checking if the value was null, but instead, we had to use our real-world understanding of the pollutants to account for any bad data.

If we wanted to be as precise as possible without bringing in so much of the human error in analysis., if we could have used an outlier technique, the cleaning data would be acceptable. Instead of simply saying 300 was too high for Sulfur Dioxide, we could instead run an algorithm against all other Sulfur Dioxide values found that a value like 300 was too high. If we were to implement this in a more professional and wanted to ensure our results were accurate, detecting outliers would be a more mathematically sound using some algorithm. After cleaning up a lot of the bad data were completely removed from our data sets as. However, there is no need to keep all attributes that would not be used for some analysis.

After we finished a lot of the data cleaning, we began to integrate the data. For much of the pollutant data, there were some different values at a different time at one day. In most instances, for things like daily max value or min value. We decided to average the values for any data on the same day. This obviously solved the issue of data redundancy in terms of having multiple values for a single day, but it also gave us a more accurate data set.

On this project first part, we created four visualizations: the scatter plot of "The NO2 AQI Mean in All U.S. States Each Year From 2000 to 2016", "The O3 AQI Mean in All U.S. States Each Year From 2000 to 2016", "The SO2 AQI Mean in All U.S. States Each Year From 2000 to 2016", and "The CO AQI Mean in All U.S. States Each Year From 2000 to 2016". For "AQI", this is "an index for reporting daily air quality"[12]. The second part, we created a line chart for Colorado AQI mean value change trend from 2000 to 2016. The third part, we created four bar chart that are four major pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) 1st max value over time (from 2000 to 2016). 1st Max Value is the maximum value obtained for four major pollutants concentration in a given day.

The design elements we used from our visualizations are "color", "shape", "space" and "tooltip". The reason is that when user put mouse on each dot, we created on the visualization, the size of the dot will change bigger, and the color of the dot or bar will change to green. The reason we do these because this will highlight the points or bars people wants to see. Also, as the year grow, the dots are at the place farther to the origin. The space is change so that people can get the direct position on the AQI value at a specific year. This is easy to people to understand the visualization. The tooltip could show some details for the dots or bar. That could help user get more information.

4. Discussion

On this project, first we created four visualizations: the scatter plot of "The NO2 AQI Mean in All U.S. States Each Year From 2000 to 2016", "The O3 AQI Mean in All U.S. States Each Year From 2000 to 2016", "The SO2 AQI Mean in All U.S. States Each Year From 2000 to 2016", and "The CO AQI Mean in All U.S. States Each Year From 2000 to 2016". For "AQI", this is "an index for reporting daily air quality"[12]. From these four visualizations, we can see that for NO2, SO2, CO, each of them has the decreasing trend, which provides an information that people are trying to control the pollutants. However, for O3, the trend is almost flat. Thus, I checked Google, it shows that "O3 is created by chemical reactions between oxides of nitrogen (NOx) and volatile organic compounds (VOC) in the presence of sunlight. Emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapors, and chemical solvents are some of the major sources of NOx and VOC"[13]. This is hard to control by human because sunlight is a major source. From my point of view, I think this is the major reason why the trend of O3 is almost flat.

Second, we focus on Colorado AQI change trend. the Carbon Monoxide AQI shows decreasing trend from the year 2000 to 2013. There are minor fluctuations during the period from the year 2001 to 2004, and the period from the year 2009 to 2013, but the main trend is decreasing. The reason we say that the main trend is decreasing is that each time the fluctuation is over, the ending AQI is lower than the AQI before this fluctuation. That is, as we can see from the graph, the Colorado Carbon Monoxide AQI at the year 2004, which is after the fluctuation, is lower than that at the year 2001, that is the time before the fluctuation. The turning point for Colorado is the year 2013. After the year 2013, its Carbon Monoxide AQI goes back up. Till 2016, the AQI has gone back to the same value as the year 2008. That is the pollution that takes five years (2008-2013) to go down only uses three years (2013-2016) to go back up.

Colorado's Nitrogen Dioxide AQI has its peak at the year 2001 and then fluctuates between the 35 and 45 until the year 2013, at which the index goes down even more and reaches its lowest value. But this value only the last a year, it goes back up the next year at a faster rate. The AQI at the year 2016 is close to the highest value of this documented period as we can see from the graph.

Colorado's Sulfur Dioxide AQI gradually becomes smaller over the years. The value at the beginning several years between 20 and 25, which a little higher. And at the end of this record, Colorado has dropped their Sulfur Dioxide AQI to values between 5 and 10, that is 10 to 15unit drop compared to 16 years ago.

Colorado has the most volatile Ozone AQI as we can tell from the graph. It has been inconsistently increasing from the year 2000 and 2015, which means the overall trend is going up but there are some backward. It has the best Ozone condition at the beginning of this record but then it becomes highest after the year 2008. And between the year 2015 and year 2016, it dramatically drops from 39 to 31, which is close to the beginning point.

Third, we focus on four major pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) max value over time (from 2000 to 2016). 1st Max Value is the maximum value obtained for four major pollutants concentration in a given day. The monitoring and analysis of ambient air pollutant concentrations are important aspects of the issue of air quality management. The evaluation of air quality is important both in terms of the intrinsic interest in the levels and variations of pollutant concentrations in the ambient air and in determining compliance with air quality standards. The ambient air quality standards for several pollutants are stated in terms of maximum concentrations not to be exceeded more than once per year. And, in the case of air contaminants, adverse effects depend on the concentration level, as well as the length of time elevated concentrations may persist. Thus, the extreme values of air quality, e.g. the maximum concentration, are of considerable interest[14].

Through our plot, we can see the four major pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) max value decrease from 2000 to 2016, just Nitrogen Dioxide and Carbon Monoxide a little increase at 2016. That means in the recent years U.S. air much better than before. This information makes complete sense logically. Air pollution has decreased even though population and the number of cars on the roads have increased. The shift is the result of regulations, technology improvements, and economic changes. This result also can answer our one question in this project we sought. We can easily predict the general trend in the future years the U.S. air condition will be better year by year.

4. 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] U.S. Pollution Data Pollution in the U.S. since 2000 https://www.kaggle.com/sogun3/uspollution

[3] 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

[4] Air pollutants measured in Seoul https://www.kaggle.com/jihyeseo/seoulairreport

[5] Dr Gerard Hoek, Bert Brunekreef, Sandra Goldbohm, Paul Fischer, and Piet A van den Brandt. 19 October 2002. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study, from https://www.sciencedirect.com/science/article/pii/S0140673602112803

[6] Dr Gerard Hoek, Bert Brunekreef, Sandra Goldbohm, Paul Fischer, and Piet A van den Brandt. 19 October 2002. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study, from https://www.sciencedirect.com/science/article/pii/S0140673602112803

[7] Shiri Avnery, Denise L. Mauzerall, Junfeng Liu, and Larry W. Horowitz. April 2011. Global crop yield reductions due to surface ozone exposure: 2. Year 2040 potential crop production losses and economic damage under two scenarios of O3 pollution,from https://www.sciencedirect.com/science/article/pii/S1352231011000070

[8] Ambient air pollution: a global assessment of exposure and burden of disease WHO Libray Cataloguing-in-Publication Data

[9] Muhayatun Santoso, Diah Dwiana Lestiani & Philip K. Hopke (2013) Atmospheric black carbon in PM2.5 in Indonesian cities, Journal of the Air & Waste Management Association, 63:9, 1022-1025, DOI

[10] Air Quality Improvement and Prospects of Shenzhen City (2016). Fusan, Korea. Human Settlements and Environmental Commission of Shenzhen Municipality.

[11] Chen, Y. S., Sheen, P. C., Chen, E. R., Liu, Y. K., Wu, T. N., & Yang, C. Y. (2004). Effects of Asian dust storm events on daily mortality in Taipei, Taiwan. Environmental research, 95(2), 151-155; Zeldin, W. (2017). "Taiwan: Action Plan to Reduce Air pollution, Shift to Electric Vehicles." Library of Congress. Retrieved from http://www.loc.gov/law/foreign-news/article/taiwan-action-plan-to-reduce-air-pollution-shift-to-electric-vehicles/

[12] Air Quality Index (AQI) Basics https://airnow.gov/index.cfm?action=aqibasics.aqi

[13] Ozone Pollution https://www.epa.gov/ozone-pollution

[14] Estimating the maximum value of autocorrelated air quality measurements Dept. of Chemical and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY 12181, U.S.A. https://ac.elscdn.com/0004698182903419/1s2.0-0004698182903419- main.pdf?_tid=504e1efe-a95a-4fe7-b7a2-5aa4b2ebab77&acdnat=1524776109_93b6e0ed4a3a5af25b6a918a1196a232