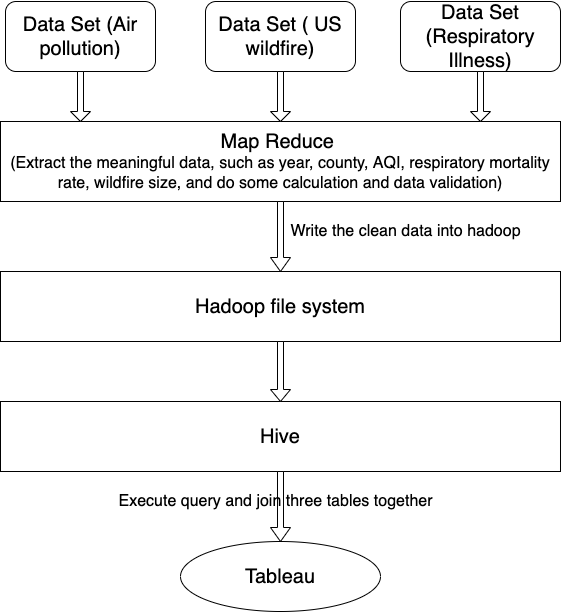
**Are wildfire, air pollution, and respiratory illness correlated in the U.S.?**

Wayne Zhu (jz3395), Dennis Shen (ts4071)

**Abstract**

Using big data in Gigabytes, we investigated the correlation between wildfire and air pollution, and the correlation between air pollution and chronic respiratory illness mortality rate across U.S. counties over the past 34 years. We attained data from mostly government agencies like EPA, cleaned them, and uploaded them to HDFS and Hive. We plotted our yearly result by counties and year on Tableau and found there existed no correlation between wildfire, air pollution, and chronic respiratory illness mortality rate across U.S. counties. Here is our overall design diagram.

****

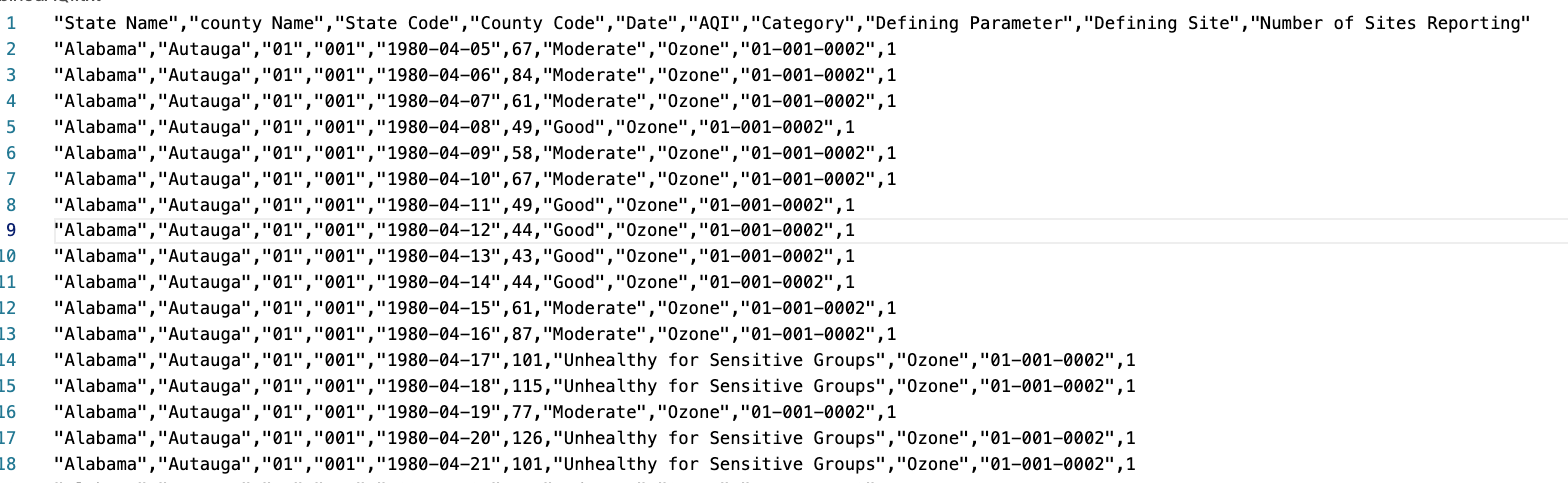
(The Data flow)

**Introduction**

Wildfire and air pollution are common tropes in the news. But are they correlated? Also, is air pollution in developed countries like the U.S. correlated with a significant increase in chronic respiratory illness mortality rate? To answer this question, we first need to ponder what kind of data we have. Using Google Dataset Search, we found US Environmental Protection Agency’s(EPA) Air Quality Index data, US Department of Agriculture's wildfire data, and Global Health Data Exchange’s chronic respiratory Illness mortality data. Upon carefully inspecting the data, we have decided to choose years as the time unit for our analysis, and the county as our geographical boundary for a big picture.

All three sets of data were either CSV or SQLite, which can be easily converted into CSV. Each dataset contained the state name, county name, and FIPS code (unique for each county). Geographically, we have decided to use only the FIPS code because we are afraid that there might be inconsistency and repetition using only names.

Then, we needed to think about what time units we use. While US chronic respiratory illness mortality data reports each county’s mortality, the wildfire and AQI data reports every day’s situation. So we wrote two MapReduce programs to calculate the yearly average AQI for each county as well as the yearly estimated areas burned by wildfire. We then uploaded all the datasets into HDFS and subsequently onto Hive so we could join the tables by the FIPScode+Year key. Now, we are ready for formal analytics.

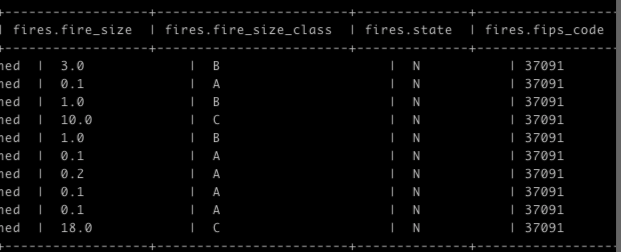


(This is the AQI data, for each row, you can see the county name, date, and daily AQI.

For our analysis, we aggregate them to get the yearly average AQI for each county.)



(Here, each row represents each county and its disease mortality rate by year. We pick the rows that contain aggregate mortality for both genders each year in each county instead of just male or female.)



(Each row represents a wildfire occurrence in each county. We get rid of some columns because this table is too large. But you can see each row has the fips county code, the fire size estimate, and the date. We add up the total acre of wildfire size per year per county for our analysis.)

**Methods and Challenges**

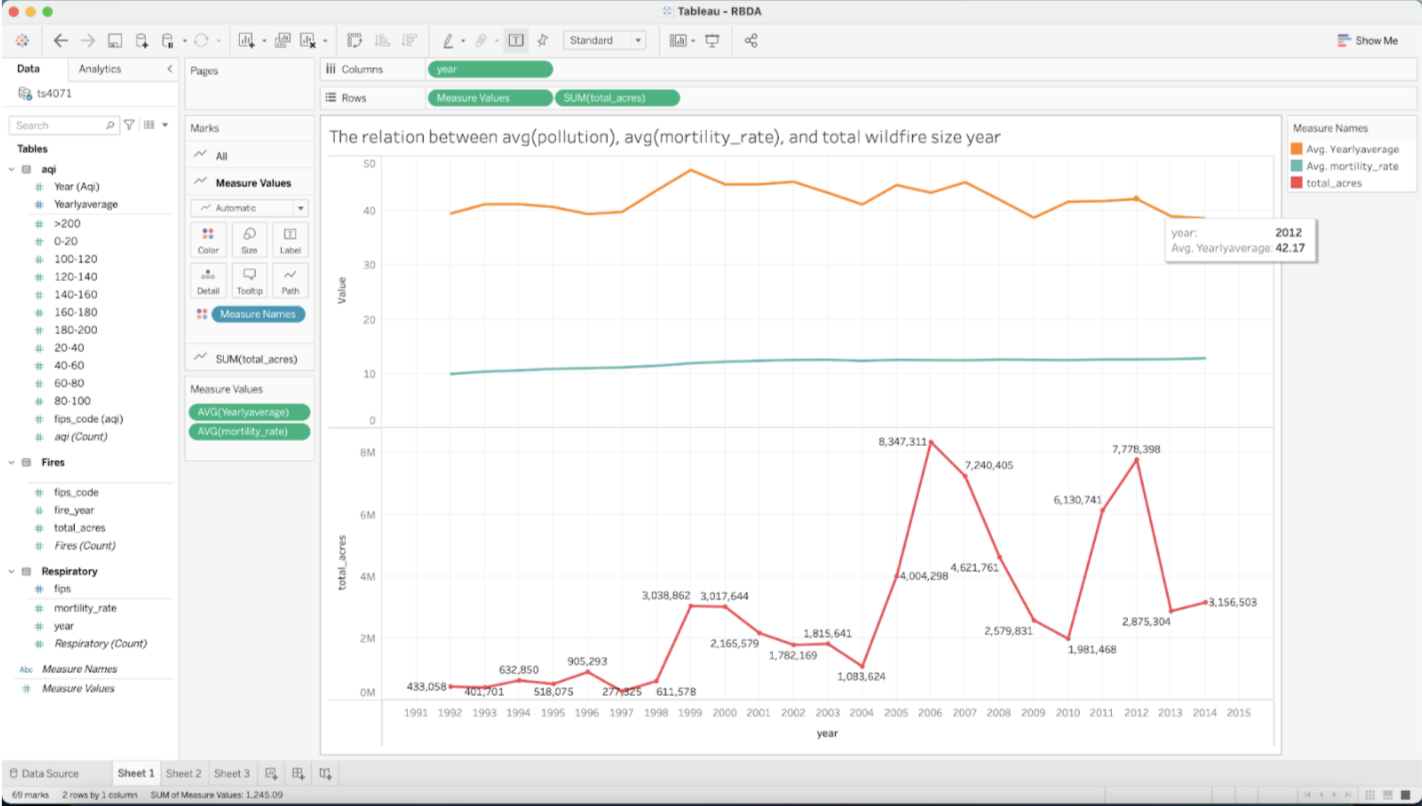
We use Tableau because we want to visualize our results. Firstly, we loaded three different datasets, pollutant, wildfire, and respiratory illness into Tableau. Secondly, we made these three tables join together by county and year. Then, we created three different graphs.

Obstacles:

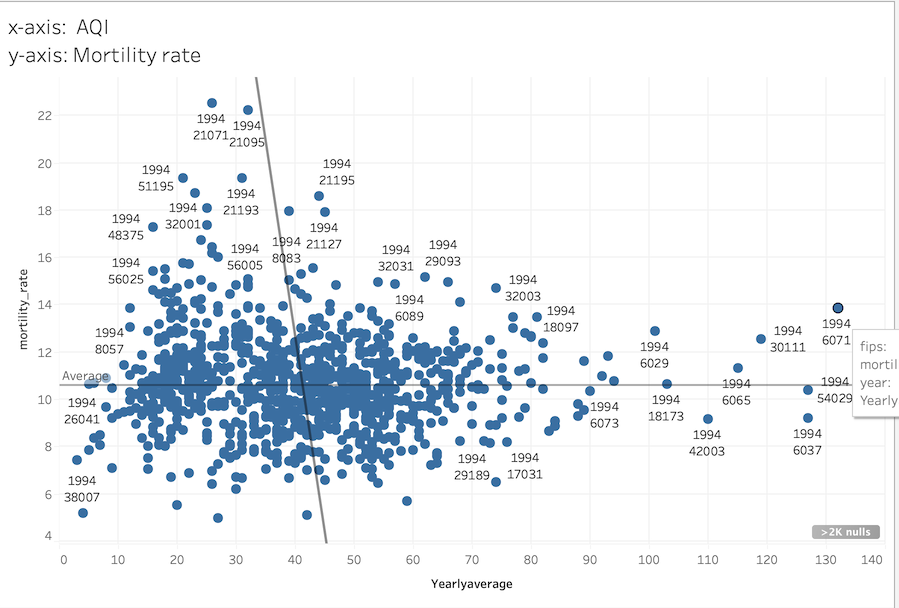
* The first challenge is that Tableau connects to the Hive database and executes the query every time. So it takes a long time especially if the dataset is big. However, we solve this problem by saving the data to our local disk but this can only work if the dataset is still not too big to save on local or the data is not continuously updated.
* Some columns have null or weird value. When we write the mapper task, it’s not easy to know the value range for each column because the dataset is saved in a CSV file. Once we put the data in HIVE, we can execute a distinct query to see all the different values. So, we filtered null or some weird values by adding a filter when we executed the query in Tableau.

**Results**

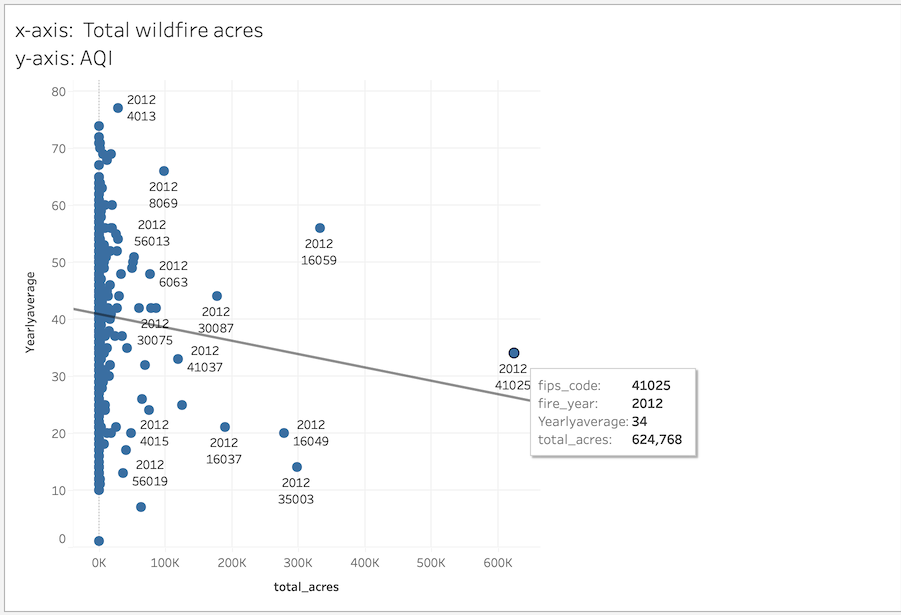
For our result, we output 3 types of graphs. The first kind shows the general trend between the average air pollution, wildfire of all counties, and chronic respiratory illness mortality over the years. As you can see, for the past 22 years, the yearly average pollution shows a downward trend but stays mostly constant. Chronic respiratory illness mortality goes up moderately and steadily. Wildfire goes up and down dramatically but has overall increased significantly. The numeric value on the y axis here is not really revealing, the trend is what matters.



Now, let’s look at the yearly data on air pollution and chronic respiratory disease mortality. We did regression analysis over 30 years, but they all looked rather similar to this one graph of the year 1994. (If you want to see all the graphs, click this [linkhttps://drive.google.com/drive/folders/1jD6SnBoIGfiiQBeyp5eFONCFTXFE5WC6?usp=sharing](https://drive.google.com/drive/folders/1jD6SnBoIGfiiQBeyp5eFONCFTXFE5WC6?usp=sharing)) The x-axis is the average AQI of each county of each year, and the y-axis is the chronic respiratory illness mortality rate of that particular year. Each dot represents a county here. Some of the years are slightly positive, others are slightly negative, but most are flat. They all show statistically insignificant correlation as illustrated by this result from 1994.



Then, we plot the yearly wildfire and air pollution, and they show no correlation also. (You can find them in the same folder here as above [link](https://drive.google.com/drive/folders/1jD6SnBoIGfiiQBeyp5eFONCFTXFE5WC6?usp=sharing)) Again, some years are negative, others are positive and most are not significant. Most of the counties do not have many wildfires at all, so outliers determine the slope. Below is a sample image for the year 2012.



**Discussion:**

There are many reasons why there are no correlation between AQI and chronic respiratory disease mortality rate.

* AQI between 0 and 50 are considered good, while AQI between 50 and 100 are considered moderate by EPA. In developed countries like the U.S., most counties are either good or moderate. According to EPA’s guideline, good air quality has no effect on human life. Moderate AQI only affects unusually sensitive people though not significantly.[[1]](#footnote-0) Therefore, it’s entirely possible that air quality is simply too good to kill people.
* Counties that score significantly high on pollutions in the 1990s like LA and San Bernadine also have high per capita income. People with higher income might have better access to healthcare or are more likely to take preventive measures like buying air purifiers. We didn’t control GDP per capita yet, so that could be a confounding variable. Likewise, there could be many other confounding variables that affect chronic respiratory illness mortality rate such as smoking that are not controlled.
* The damage of air pollution, especially PM2.5, is cumulative. The person only falls ill after enough dust accumulate in his lung. It’s possible that pollution in one year might cause people to fall sick later. This hypothesis of delayed effect is supported by our over the years trend graph—as air pollution comes down gradually, the respiratory illness mortality goes up.
* It’s also possible that bad air only weakens your lung and make you more vulnerable to other acute respiratory illness. But your death cause won’t be chronic respiratory illness itself—it could be, for example, covid instead.

The lack of correlation between wildfire and air pollution also makes sense because wildfire typically only lasts for a few days. So it won’t affect the overall air quality of the area in the entire year. Also, the pollutant generated by wildfire could easily be carried to other counties. For example, I still remember earlier this year when wildfire pollutants from California flew all the way to New York City and blew AQI up to above 100. This will be a serious confounding factor since the smog travelled through virtually the entire country horizontally.

**Date source References:**

* The US EPA air pollution data as found here. All the years from 1980-2021

<https://aqs.epa.gov/aqsweb/airdata/download_files.html>

[Download Files | AirData | US EPA](https://aqs.epa.gov/aqsweb/airdata/download_files.html)

* Spatial wildfire occurrence data for the United States, 1992, 2018

<https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0009.5> (905MB)

* The US Chronic Respiratory Illness Mortality rate by County from 1980 to 2014

<http://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014>

|  |
| --- |

1. https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf [↑](#footnote-ref-0)