**COVID-19 & Air Quality**

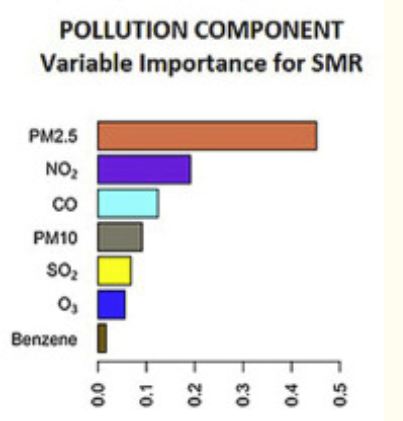
*The Wranglers*

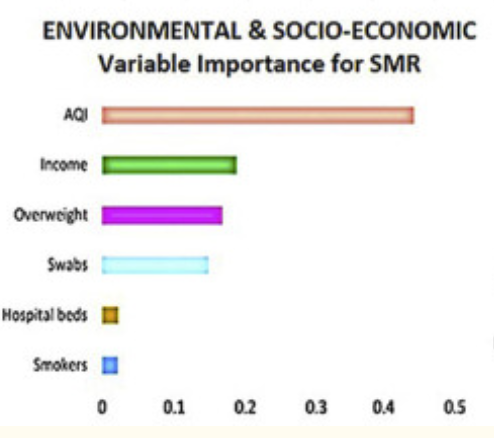
*Adriana Reyes Miranda, Teigen Judd, Jonathan Barton, Yi-Jin Chen*

**Original study**

**Problem**

The original study used machine learning methods to reveal the prolonged exposure to air pollution associated with SARS-CoV-2 in Italy. They explored environmental and socio-economic factors for SARS-CoV-2 standardized mortality ratio (SMR) and analyzed the pollution sector and pollution component. In their results, the top three pollution components were: PM2.5, NO2 and CO. PM2.5 is the main component and the most crucial predictor of SARS-CoV-2 effects even with a slight decrease of air quality (AQI).





**Differences between the United States and Italy**

The US has a much larger area than Italy, meaning there is possibility for a much higher variance in air quality data. And though the US has a higher population, Italy has a more dense population. Most of the US is rural.

**Original Description of the data**

**Air Quality Data**

The air quality data comes from EPA (<https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html>). The EPA keeps measurements of various air pollutants at thousands of locations across the United States. This data is publicly available at EPA.gov. In the beginning, we downloaded data from Outdoor Air Quality Data (<https://www.epa.gov/outdoor-air-quality-data/download-daily-data>). Each pollutant had to be downloaded separately, broken down by year and geographic area. Since that process is time consuming, we decided to use a different website to get all the required data easily (<https://aqs.epa.gov/aqsweb/airdata/download_files.html>,thanks to Professor Naomi Riches for sharing.)

**COVID-19 Data**

COVID-19 data from the Johns Hopkins Center from Systems Science and Engineering repository:<https://github.com/CSSEGISandData> . This data is presented in a visual dashboard format that draws from multiple sources which include: WHO, CDC, European Centre for Disease Prevention and Control (ECDC), and several other institutions. Data is gathered for each US state from sources often provided by the state, city, or county. Data is also collected from several independent national sources. The data is licensed under Creative Commons for public use. At the request of the data providers, credit is given to: “Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1” as well as: “ Badr, H. S., B. F. Zaitchik, G. H. Kerr, J. M. Colston, P. Hinson, Y. Chen, N. H. Nguyen, M. Kosek, H. Du, E. Dong, M. Marshall, K. Nixon, and L. M. Gardner, 2021: Unified COVID-19 Dataset.” The data is available for public viewing in both desktop and mobile formats. Two data points (COVID cases and COVID deaths) are tracked at various levels of organization.

**Data quality report (original data)**

**Air Quality Data**

Pollutant Selection and Data Profiling:

After reading the paper we decided to work with AQI and selected the top

3 pollutants used in the paper, which were: PM2.5, NO2, and CO.

We first used California as an example to check the data, but the instructor later told us that we should strive to make this work more representative of the whole US population and used the 50 states. We obtained the amount of counties per state for original pollutants that we want to work with. After checking the whole nation with county data, we found that not all states had CO data for more than 50% of their counties. With this, we decided to use Ozone instead of CO.

Unreasonable data:

We observed that there were negative values for pollutant concentrations and AQI. The data also had some states listed in the “State Name” column that are not actually states (like DC, Puerto Rico, and the Virgin Islands).

End-date Mismatch:

We noticed in the original files, ozone only recorded to Nov 14, NO2 and PM2.5 are both recorded to Oct 31.

**COVID-19 Data**

The COVID data presented some interesting challenges. From our initial profiling and analysis, we noticed that many states began reporting COVID data on different dates. We found that while the dataset was mostly comprehensive, there were some oddities in reporting. For example, New York City was reported all together, even though New York has several distinct boroughs in the air quality dataset. When profiling, we also discovered that the data for case and death counts was a running total, so we had to perform functions that calculated daily cases and deaths from the aggregated totals.

**Wrangling Steps**

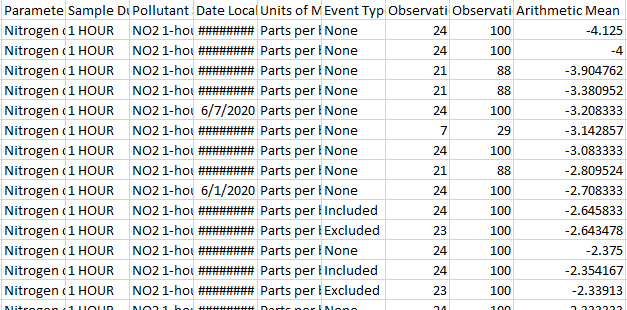
**Air Quality Data**

Changing selected pollutants for analysis:

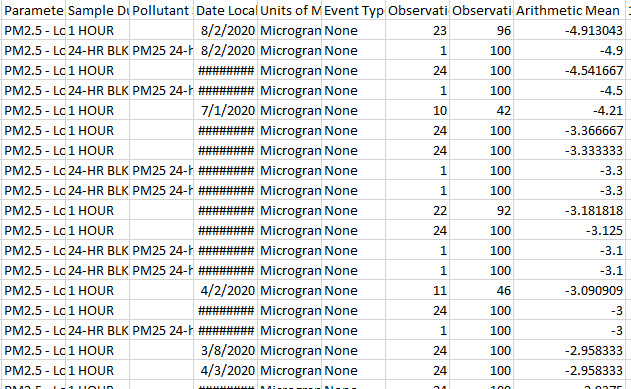
We profiled PM2.5, NO2, and CO data separately from daily summary data with all states and counties. We also checked the whole nation with county data of each pollutant. After profiling and discussion, we decided to keep PM2.5 and NO2, but swap CO for Ozone.

Clean:

During profiling, we found there were negative values for pollutant concentrations and AQI. For example, in NO2 “Arithmetic Mean” and PM2.5 “Arithmetic Mean,” both had negative values shown in the following screenshots. After discussing with Ram, we decided to change negative sensor values to 0 and profiled the data again. After running boxplots on the CO data, we discovered several outliers. We thought about replacing them with median, but after discussion with the professor, we decided to keep these outliers as we don’t really know whether these are true outliers or not.



**NO2 screenshot**



**PM2.5 screenshot**

Visualization issue:

After profiling Ozone data, we found that the largest “Arithmetic Mean” value is 0.121 Parts per million (ppm). This small value is not easy to be compared against AQI, which has a largest value of 266. Even when comparing with NO2, the largest “Arithmetic Mean” value is around 60 parts per billion (ppb), and the largest “Arithmetic Mean” value is about 577 Micrograms/cubic meter (LC) in PM2.5. Therefore, for visualizations and dealing with pollutants’ different units, we decided to derive a ppb ozone arithmetic mean variable. In this way, at least ozone and PM2.5 have the same unit, and it is easier to tell from visualization graphs.

Data merging:

After profiling and cleaning each pollutant’s data, we selected the time series we wanted (from May to December 2020). We selected the columns we wanted to analyze, and those columns are “Date Local,” “State Name,” “County Name,” cleaned Arithmetic Mean,” and “AQI.” We combined Ozone, NO2, and PM2.5 to a completed air dataset, so we knew which dates, states, and counties could be used to match with COVID-19 data.

**COVID-19 Data**

We profiled the COVID data to determine how many null values were present. After getting an idea of data completeness, we used the interpolate function to fill null values with the nearest value from the same county. This provided the most accurate data possible. Following the use of the interpolate function, all remaining null values were set equal to zero and negative values were also set to zero. It is hypothesized that some of these negative values existed due to mis-reporting of data.

Because the air quality dataset was the limiting dataset (fewer locations recorded than COVID data), it was chosen as the model dataset. After this data was cleaned, we had to ensure that we could match all locations and dates in the AQ data with COVID data. To do this, we made a unique merge ID consisting of state, county, and date. This merge ID was added to each row of both the COVID and AQ data. Initial compatibility was measured by making a set of unique merge IDs for both datasets, and then subtracting unique IDs of the COVID data from the AQ data. This gave us a list of values that were in the AQ dataset, but not in the COVID data. To fix this, data points were changed via python to ensure a smooth merge. Below are some examples of changes that were made to the COVID dataset:

* + - 1. Change “St. Clair, Illinois” to “Saint Clair, Illinois” to match with AQ data merge ID
      2. Change “Hampton city, Virginia” to “Hampton City, Virginia” (capitalized ‘H’) to match with AQ data merge ID
      3. Change “Denali Borough, Alaska” to “Denali, Alaska” to match with AQ data merge ID

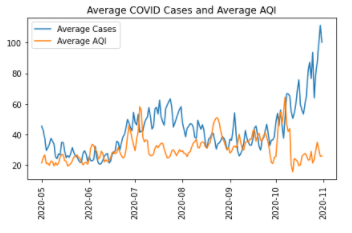
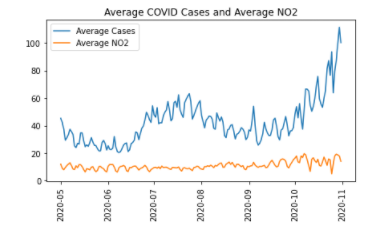
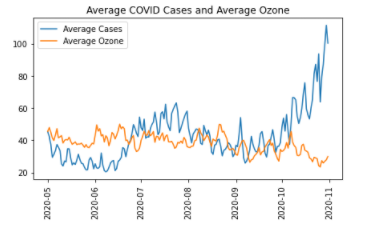
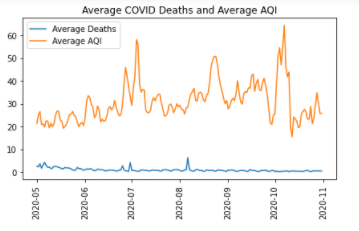
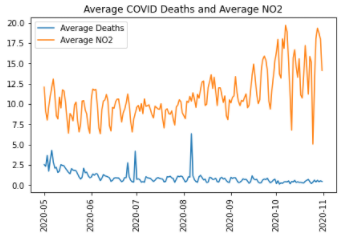
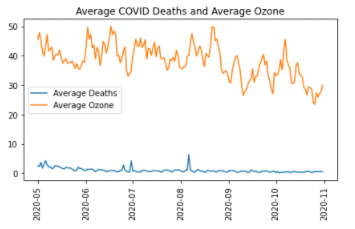
Following these changes to the data, we performed another subtraction to see which merge IDs were not present in both datasets. Once we got this list, we were able to look into each individually and determine why they were discrepant. (see below) These discrepancies were all due to delayed COVID data reporting. Each row for these counties will show AQ data, but will have null COVID data until the data reporting start point.

* + - 1. Custer, South Dakota started data reporting on 6/4/2020
      2. Rosebud, Montana started on data reporting 6/2/2020
      3. Avery, North Carolina started on data reporting 5/19/2020
      4. Butte, Idaho started on data reporting 8/11/2020
      5. Phillips, Montana started on data reporting 8/5/2020
      6. Taylor, Wisconsin started on data reporting 5/21/2020
      7. Denali , Alaska started on data reporting 6/25/2020 (Denali Borough)
      8. Fergus, Montana started on data reporting 6/18/2020
      9. Powder River, Montana started on data reporting 7/23/2020
      10. Bell, Kentucky started on data reporting 5/18/2021
      11. Billings, North Dakota started on data reporting 6/20/2020
      12. Palo alto, Iowa started on data reporting 5/8/2020
      13. Jackson, SD started on data reporting 5/24/2020
      14. Trego, Kansas started on data reporting 5/22/2021
      15. Monroe, Miss started on data reporting 5/21/2020
      16. Weston, Wyoming started on data reporting 5/27/2020
      17. Mercer, North Dakota started on data reporting 5/5/2020
      18. Yancey, NC started on data reporting 5/3/2020
      19. Lake, Minnesota started on data reporting 5/7/2020
      20. Forest, Wisconsin started on data reporting 5/6/2020

**Report of the quality of merged data**

* 1. These following graphs show proof of successful merge and give a brief view of possible relationships between data (across the entire US).

Line graph

* + - 1. 
      2. 
      3. 
      4. 
      5. 
      6. 
  1. Linear regression

