CSC110 Project Report: COVID-19 In the Air

Helia Sajjadian Moosavi, Ipek Akyol, Yumna Refai

Monday, December 13th, 2021

1. Problem Description and Research Question

Research Question: How has the impact of COVID-19 on different sources of emissions increased or decreased air pollution in China?

During the pandemic, one of the main topics of discussion was how government imposed lockdowns that aimed to reduce the spread of COVID-19 led to a halt in the production of goods and services in various industries. This inadvertently lowered air pollution and contributed to the reducing of climate change in some of the most polluted countries in the world, particularly China. Considering that China had some of the strictest lockdowns and was one of the most polluted countries in the world pre lockdown, finding data on pollutant concentrations in China was particularly interesting to our group.

Inspired by the 2021 UN Climate Change Conference that is currently taking place, we found climate change to be a relevant topic related to COVID-19's impact on the world. Since climate change has had a significant impact on not only the environment, but also on community health, we have been interested in analyzing the impact of the pandemic on air pollution.

Within our project, we analyzed how COVID-19 has affected air pollution while specifically focusing on the different sources of emission as well as the corresponding concentrations of pollutants emitted daily. Therefore, we found it best to look at data about pollution concentrations in China and create a model to show whether air pollution changed due to COVID between 2019 - 2020.To assess whether air pollution has increased or decreased with the impact of COVID-19, we first furthered our research with the following questions:

1. Which sources of emission had the least percent change in pollutant emissions during COVID-19?
2. Which sources of emission were impacted the most by COVID-19?
3. Which industries within sources of emissions were affected the most by COVID-19?
4. For each industry within sources, how did the emission of pollutants increase or decrease during lockdown?

Based on the feedback received by the TA, we also decided to compute a linear regression model to see if Covid19 was a predictor of the change in total emissions of pollutants by each source. Here, Covid19 is indicated by the time frame: from Jan 2020 to March 2020.

2. Dataset Description

Our original data set named 'Data_20210712.xlsx' was obtained from a reliable and free online platform named 'Mendeley Data'. It was designed to provide real-life statistics on daily emissions of air pollutants like CO_2 in the Yangtze River Delta region of China. The region is made up of the Chinese Speaking areas of Shanghai, southern Jiangsu province and northern Zhejiang province. The dataset specifies the time-frame of before and during COVID between 2019 (January - March) and 2020 (January - March) respectively for this region.

The format of the dataset is xlsx. The data given includes information on different sources of emissions, the smaller industries within these sources, and corresponding pollutants for each date with the distinction of before and during Covid19. Within the dataset;

- There are 5 different **main sources of emissions**: Power Plants, Heavy Industry, Light Industry, Mobile Sources, and Other Sources.

- Each main source of emissions is broken down into narrower categories, which we will call **industries**. For example, the industries under mobile sources are Gasoline Vehicles, Diesel Vehicles, Non-road Machinery, Marine, and Aviation Aircraft.

- There are **8 pollutants**: SO_2, NOx, CO, NMVOCS, $PM_{2.5}$, BC and CO_2. For each pollutant type, the same sources of emission and industries are listed. Observations within the dataset show the concentration of each pollutant emitted within each industry. These observations correspond to the rows. As an example, we have ”Mobile Sources” listed 8 times under the source of emission column, each corresponding to an emission value of a different pollutant.

- The initial original dataset was then renamed to ’data.xlsx’ which contained the same rows and columns of the original dataset, except the previous first two rows are now un-merged. We then used all the columns except for columns D and E which corresponds to 2019 base and 2020 base to address the questions specified in the introduction.

3. Computational Overview

The purpose of our project is to highlight the impact of Covid on different sources of emission and it’s industries as well as point towards the increase and decrease in emissions throughout the COVID-19 period.

A. Data Transformation and Aggregation In order to achieve this purpose, one computational method that we used is data transformation and aggregation. During this step, we organized our data utilizing a new Python library openpyxl. This library enabled us to read the collection of data found within our main data set (provided in Excel format) and transform it into readable and organized pieces of information by Python.

Then, we created three data classes: data class that represents the days within our chosen time-period where the concentration of pollutants emitted for each main source and industry for each pollutant is recorded daily, one data class that represents the pollutants that contribute to air pollution and another one data class that represents emission sources. We named these three data classes Day, Pollutant, and Source.

- The Day class has the attributes date (the datetime value of the date that the observations were recorded), total (the float value corresponding to the total concentration of pollutants for that day), pollutants (dictionary with keys of names of pollutants emitted for that day and values that correspond to the objects of the Pollutant class).

- The Pollutant class has the attributes name, total (float value corresponding to the sum of the concentrations of pollutant emitted by each main source), and sources (dictionary with keys as the name of sources that emit the pollutant and values as the object for each source).

- The Source class has the attributes name, total (integer value corresponding to the sum of the concentrations of pollutant emitted by each industry within the source for a specific pollutant), and industries (dictionary with keys as the name of each industry and values as the float value of concentration of pollutant emitted by that industry).

```
class Day:

    date: datetime
    total: float
    pollutants: dict[str, Pollutant]
```

```
class Pollutant:

    name: str
    total: float
    sources: dict[str, Source]
```

```
class Source:

    name: str
    total: int
    industries: dict[str, float]
```

Finally, we converted our xlsx file into readable code which was then organised into a list containing the type ’Day’ which condenses the original dataset.

B. Computational Models and Visualization To find answers for the four main questions that we included within the Problem Description, we explored new libraries such as pandas, NumPy, math.plotlib and scikit-learn.

- **Question 1**: Which sources of emission had the least percent change in pollutant emissions during COVID-19?
- **Question 2**: Which sources of emission were impacted the most by COVID-19?
- **Question 3**: Which industries within sources of emissions were affected the most by COVID-19?
- **Question 4**: For each industry within sources, how did the emission of pollutants increase or decrease during lockdown?

In our visualisations file, we utilised the libraries - matplotlib, numpy, panda and sklearn and it's corresponding in built methods to create pie charts, bar graphs and linear models.
We mainly used mathplotlib which was abbreviated to plt by us to plot both pie graphs and bar plots with data. We also used this library to create a title and edit the legend. Plt was also used to change the x, y axis labels, and fontsizes of titles, labels and keys in the legend. We also enabled the pie chart to display the percentage decrease and this was possible because this was one of the inputs specific to only plot.pie from mathplotlib.

To obtain the data used for plotting from our dataset, we imported the get_data which stores the different subsets of the data. From within the data, we filtered the total pollutant emissions emitted daily by sources of emissions before Covid (2019) and after Covid(2020) as well as those emitted daily by the industries within the respective sources before Covid (2019) and after Covid(2020).
C. Linear Regression Model For the final part of the project, we utilized linear regression models to make inferences about COVID-19's impact on total pollutant concentration. We used the pandas, NumPy, and math.plotlib libraries to plot the graph of the linear regression model and aimed to use the scikit-learn library to fit the regression equation.

Linear Regression assumes that there is a best straight line that explains the real relationship between two variables x and y and that the values we observe within our original data randomly deviate from this straight line. The equation of a simple linear regression model for an observation i is:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

In this equation, y is the response/dependent variable for observation i and x is the predictor/independent variable for observation i. β_0 represented the intercept of the linear regression line and β_1 represented the slope parameter. ϵ_i is the random error term for observation i.

For our linear regression model:

- We determined x as the number of days that passed from the first date within our data set (for example, the first day has an x-value of 0). We determined y to be the total pollutant concentration (sum of the pollutant emission of all industries) for that day.
- Then, we aimed to get an estimate for β_0 and β_1. This process would provide us with an estimated simple linear regression line with intercept $\hat{\beta_0}$ and slope of $\hat{\beta_1}$:

  $$\hat{y} = \hat{\beta_0} + \hat{\beta_1} \cdot x$$

  $\hat{y}$ is the estimated average value of total pollutant concentration on a certain day. $\hat{\beta_0}$ is the average of total pollutant concentration for day 0. $\hat{\beta_1}$ is the average change in total pollutant concentration for 1-unit change in x, which corresponds to a one day difference.
- From this equation, we hoped to be able to calculate on average how higher or lower the total pollutant concentration is on x days after the first day of observations.

After plotting the graph of total pollutant concentration vs. number of days for each pollutant, we discovered that the association between these two variables were not linear as we hypothesized. Therefore, we could not fit the regression line to get the estimate values we needed. However, this realization allowed us to come to conclusions regarding our main research question. These details are discussed further in the discussion section of this report.

4. Instructions for Obtaining Datasets

- First, save the dataset as 'data.xlsx' in the folder along with the other python files in one source directory. Mark the directory as Source Root. So the files you will have apart from our latex file and pdf are :-
  1. main.py
  2. get_data.py
  3. visualisation_functions.py
  4. data.xlsx
  5. Data_20210712.xlsx
  6. requirement.txt

- Then, install all the libraries mentioned in the requirement.txt file.
- After this, open the get_data.py and make sure you have data.xlsx in your directory. Run the file. You can convert the xlsx file using the first function and then organise the data into a condensed, interpretable list using the second function.
- Then, start getting necessary data about the sources of emissions or it's industries to build the different visualisations i.e. pie charts, bar plots and linear regression models using the functions in this python file.
- Now, go to the visualisations_functions.py file. In here, you have access to all the functions that draws either pie charts, bar plots or linear regression models. Since, we've imported data from get_data.py into this file, we can call the respective functions in the console and the respective visualisation/ model will appear.
- Now that you have tested all the functions in both files, go to the main.py which is a module that contains all of the code necessary to run the entire program from the beginning to the end.

5. Changes Between Proposal and Project

Moving from the proposal stage of our project to the actual implementation of the project, we specified the questions that we want to thoroughly explore and determined more tangible methods than suggested in the proposal.

We also included computation models to build linear regression models for each pollutant gas and following this, we built a dataframe that consisted of predicted and actual values of pollutant concentrations to portray whether our predicted values are close to the actual values and hence, indicates our model is performing well.
Instead of using plotly, we used pandas, NumPy, and math.plotlib to plot our pie charts, bar graphs, and other models. This allowed us to discover more libraries than we previously predicted.

6. Discussion

The results of our computational exploration helped us answer the four sub-questions that we indicated in our project description. The answers to these questions provided us with the inferences that we needed to come to a conclusion regarding our main research question. As discussed in our computational overview, we were able to come to several conclusions by using the visualization of our results:

- Mobile sources had the greatest percent change in pollutant emissions. This result was visualized in a pie-chart that showed the weighted decrease in pollutant emissions by each source.
- Mobile sources also had the highest absolute change in pollutant emissions from before COVID (2019) to during COVID (2020).
- All main sources had a decrease in total pollutant emissions when moving from 2019 to 2020.
- The cement manufacturing industry had one of the largest decreases in the total pollutant of gases emitted when moving from 2019 to 2020.
- Several industries such as Iron Steel and Chemical Manufacturing saw an increase in the total of all pollutant gases emitted when moving from 2019 to 2020.

- The values for the total emissions for the eight pollutants were significantly similar to each other, yet even the smallest difference in percentage values meant tons of pollutants emitted.

Utilizing this information, we came to the conclusion that COVID-19 has decreased air pollution in China overall. However, another one of our goals was to determine if there was a linear association between the number of days that passed during COVID-19 and total pollutant concentration.

Using linear regression during the computational part of the project, we had aimed to determine what kind of an impact COVID-19 had on total pollutant concentration as well as predict the value of total pollutant concentration for a certain day within the time frame of our data set.

One of the limitations that we faced regarding the data set that we have found was the quadratic nature of the graph of total concentration of pollutants vs. number of days. Our initial hypothesis was that the total pollutant concentrations would decrease in a linear fashion over the time period or simply, act inversely proportional to the number of days. However, with our computational model, we quickly discovered that there was not a linear association between the two variables.

Even though we could not fit linear regression lines for each pollutant, we discovered that there were large dips in the graph that corresponds to the decrease in total pollutant concentration around day 400. Although the significance of this time frame is not largely discovered, there being a decrease in the emission of each pollutant suggests an association between the number of days that have passed and the y-variable.

The number of days that have passed is not necessarily significant as it depends on the initial day of our time frame. However, since this time frame coincides with the beginning of COVID-19, we collect evidence towards the hypothesis that the impact of COVID-19 on different sources of emissions has decreased air pollution in China.

An extension that we could add to our research project would be to extract data from time periods where total pollutant concentration is linearly decreasing or linearly increasing. This would allow us to almost separate the quadratic graph into two different graphs. From these two graphs, we could once again predict trends within the increase and decrease of total pollutant concentrations using linear regression models.

When interpreting the results of these linear regression models, one aspect that we would need to consider is the prediction accuracy of our model. Prediction accuracy allows us to quantify how accurate the predictions of our model are. To achieve this, we have to randomly divide the sample data into training and testing data sets. 80 percent of the data is randomly selected for the training data set and 20 percent of the data is selected for the testing data set.

To discuss our results and the limitations of the linear regression model, we would have to calculate the root mean squared error. RMSE is calculated by the equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}}$$

A small value of RMSE would suggest good prediction accuracy as it shows that the predictions for the response variable ($\hat{y_i}$) are close to the observed values (y_i).

Another area of exploration is to determine the factors that triggered an increase in the total pollutant concentrations after the dips during COVID-19. Since total pollutant concentrations have began to increase as the impact of COVID-19 has slowly passed, we can infer that we need to take additional steps as individuals, corporations, and governments to contribute to the decrease in pollutant emissions and the efforts for climate change.

7. References

Gazoni, Eric, and Charlie Clark. “A Python Library to Read/WRITE EXCEL 2010 Xlsx/XLSM Files¶.” Openpyxl, https://openpyxl.readthedocs.io/en/stable/.

“GLASGOW CLIMATE CHANGE CONFERENCE.” Unfccc.int, unfccc.int/conference/glasgow-climate-change-conference-october-november-2021.

Huang, Cheng. “Covid-19 Emission Data in East China.” Mendeley Data, Mendeley Data, 13 July 2021, https://data.mendeley.com/datasets/92mp3bbxpy/1.