Group 17: Impact of COVID-19 on Air Pollution

Simei Li
Vrije Universiteit Amsterdam
s8.li@student.vu.nl

Yiran Li
Vrije Universiteit Amsterdam
y46.li@student.vu.nl

Zhining Bai
Vrije Universiteit Amsterdam
z.bai@student.vu.nl

ABSTRACT

Affected by the COVID-19 outbreak, industrial production and human social activities have been significantly reduced, leads to decreasing in polluting gases in the atmosphere. In order to visualise the data on gases affected by the coronavirus, the main target of this project is to analyse and use the data captured by Tropospheric Monitoring Instrument(TROPOMI) to create a visualisation of the air quality change processes on both global and regional scale. In the project, the final data is stored in GeoJSON files after data processing on Databricks. The data pipeline starts with a selection from raw data, data quality assessment and data cleaning, integration and calculation. After removing the data which qa value is lower than 0.5 (0.75 for NO_2) and which on the ocean and Antarctica, and averaging points with similar latitude and longitude, the data size was decreased by over 99.6%. The final product is a global visual interactive map of selected gas concentration distribution between 2019, 2020 and 2021 using Mapbox.

1. INTRODUCTION

The COVID-19 lockdown and travel restrictions worldwide did lead to the emissions drop of key air pollutants in 2020, especially in urban areas. According to the World Meteorological Organization (WMO) 's Air Quality and Climate Bulletin, though areas such as China, Europe, and North America have a reduced concentration of aerosol, many parts of the world still do not meet the aerosol concentrations of WHO guidelines. To improve the air quality in the long term should still be human's sustained goal in the future. During COVID-19, people gradually accept life with the stay-at-home and work-from-home pattern, which could significantly affect air pollution. Currently, countries are gradually emerging from lockdowns which will probably increase the air pollution worldwide. This paper will conduct visualization based on the TROPOMI data from 2019 to 2021. Therefore, in this project we visualized the air pollution dataset compared between before, during, and after COVID-19 lockdown.

Tropomi is a Dutch-made satellite instrument on board the Copernicus Sentinel-5 Precursor satellite. It can accurately survey air quality and provides the data source of this paper. The main objective of the Copernicus Sentinel-5P mission is to perform atmospheric measurements with the high Spatio-temporal resolution to be used for air quality, ozone & UV radiation, and climate monitoring & forecasting. The main pollutants, such as aerosol, sulfur dioxide (SO_2), nitrogen oxides (NO_x), carbon monoxide (CO), and ozone (O_3), are usually used for evaluating air quality. In this paper, we decide to generate visualization on NO_2, SO_2, CO, and CH_4 gas.

Some literature has shown the changes and analysis of pollutant gases from 2019 to 2020 based on TROPOMI data. Therefore, this project will add 2021 data to conduct a deeper investigation. As the final data visualization product, we decide to provide a static html. There will be a website to show the change of various types of gas concentration worldwide.

In this report, section 2 contains a literature review related to air pollution, TROPOMI data, essential considerations, and visualization. Section 3 lays out the research questions, which help to clarify the goals and organization of this project. Section 4 gives a detailed description of the data investigation, data analysis and the pipeline,as well as the visualization website and analyses the visualisation outcomes. The cloud cost is calculated in Section 5. Section 6 provides a summary of the project, also addressing to the problems raised in section 3, analyzes the challenges faced and lessons learned, and lists flaws and improvements.

2. RELATED WORK

2.1 Key Pollutant Gas

The hazards of air pollutants have a wide range of consequences. Respiratory illnesses and physiological dysfunction are the most serious threats to the human body. The impact of air pollution on weather and climate is likewise significant. Acid rain, for example, is rainwater that contains sulfuric acid. As a result, crops are ruined, and buildings are corroded and polluted.

Carbon monoxide (CO), nitrogen dioxide (NO_2) and sulphur dioxide (SO_2) are all regarded major air pollutants in cities because they are directly discharged into the atmosphere by fossil fuels burned in power plants and automobiles.

Proceedings of the VLDB Endowment, Vol. 12, No. xxx
ISSN 2150-8097.
DOI: https://doi.org/10.14778/xxxxxxx.xxxxxxx

The principal source of CO is incomplete combustion of carbon-based fuels. The transportation sector plays a key role for more than 80% of all CO emissions. When the heating system is turned on, as well as in the petroleum refining, chemical production, and other industries, CO is released [5].

NOx is formed during combustion when nitrogen and oxygen gases react in the air, especially at high temperatures. Traffic accounts for around half of the output, with the rest coming from households and industry [11].

SO_2 is mainly generated from the combustion of sulfur-containing fuels (such as coal and petroleum), the production process of chemical, oil refinery and sulfuric acid plants, as well as electricity generation from power plants.

Despite the fact that methane (CH_4) is not a hazardous chemical, it is a greenhouse gas that contributes significantly to global warming. It is generated when organic material breaks down or decays, and it can be released into the atmosphere through natural processes such as plant decay in wetlands or human operations like oil and gas extraction and waste management.

Overall, combustion activities in industry, as well as automobile traffic, are the main producers of air pollution. According to the source of these gases, human behavior will have a greater impact on those hazardous gases than on CH_4 [17].

By analyzing air pollution in terms of seasonality, we can deduce the presence of elevated levels during the heating season, as well as low levels throughout the summer season. It follows a regular pattern that air pollution was reduced when the temperature was high, the wind speed was low, and the humidity was low [3].

2.2 Impact of COVID-19 on Air Pollution

The aim of this experiment is to show how COVID-19 affects air pollution. As mentioned in Section 2.1, gas emissions are related to the frequency of human activity, and the spread of COVID-19 has prompted many governments to enact blockade policies, so the air pollution problem should diminish in 2020 when COVID-19 broke out, and polluting gas emissions should increase in 2021 when countries lift or relax their blockade policies.

Janet et al. (2020) studied gas emissions in China during the COVID-19 period of November 2019 to April 2020 [13]. Compared with the previous winter, there is an increase in aerosols over most of Northeastern and Central China. At the same time, NO_2 concentrations declined sharply. The contributors to increased atmospheric particulates may include inflated industrial production and low wind speeds. A study conducted by Mahato and Pal (2020) in India claimed that after forced restrictions on outdoor activities, the concentrations of PM2.5, NO_2 and CO levels in the outdoor air reduced during the lockdown phase, and the air pollution levels dropped rapidly which sparked the discussions of the lockdown being an effective solution to control air pollution [12].

2.3 Important considerations

Before visualizing the TROPOMI data, it is necessary to consider all factors that influence the air pollution pattern and concentration, influencing the scope of the data selection in this project. All potential sources of emission, pollutant chemistry, transport, and bias in measurement and meteorological and environmental conditions would influence the air pollution condition [2].

According to the Copernicus program, it is essential to consider the data quality related to cloud cover. Cloud obscures the accuracy of the NO_2 concentration, which will lead to flawed estimates [4]. Moreover, in the Product Readme File of NO_2, qa_value > 0.75 is a recommend pixel filter, which could remove cloud-covered scenes (cloud radiance fraction > 0.5), partially snow or ice-covered scenes, errors, and problematic retrievals [8].

Regarding to Section 2.1, the weather impact the air pollution concentration. According to Tijl Verhoelst 's research, NO_2 concentrations are highest in the winter and summer (mostly December and June) and minima near the equinoxes. There is a seasonal cycle, with the largest values observed in the winter of the Northern Hemisphere [19].

2.4 Visualization

As for all Sentinel missions, the Sentinel-5P products are freely available to users via the Copernicus Open Access Hub. In 2016, Nicolas et al. gave a thorough description of the operational TROPOMI SO_2 algorithm and the S-5P SO_2 L2 Algorithm Theoretical Basis Document v1.0 [18]. Yuping et al. (2019) analyzed the seasonal variable characteristics of the CO total column from June 2018 to May 2019 in China based on the TROPOMI dataset [10]. Their analysis of the CO curve trends by region provides some ideas for visualisation. In 2020, Lerato Shikwambana et al. used the Earth Engine Code Editor and the QGIS software to analyze SO2 and NO2 TROPOMI OFFL datasets in South Africa. The datasets period is from December 2018 to September 2019 [16]. In 2020, Bauwens et al. assessed the impact of the coronavirus outbreak on NO2 pollution using TROPOMI and OMI observations [1].The Atmospheric Toolbox was developed for the European Space Agency (ESA) by S&T Corp in partnership with EUMETSAT and the Royal Belgian Institute for Space Aeronomy. Tropomi (Sentinel-5P) is one of the support data to the toolbox. In addition, many helpful visualisation examples can be found on Github, and Mapbox Javascript API also provides the tools to create and share interactive web maps with static html.

3. RESEARCH QUESTIONS

The object of this project is to explore the impact of the COVID-19 pandemic on the content of multiple pollutant gases in the troposphere through data visualization. The raw data need to be dealt with comes from TROPOMI, which the file are in NetCDF type. In order to present the best results, the following issues were raised when performing the project design and they need to be investigated during the project design process.

- Which variables are we interested in and how to filter out the most significant data?
- What is the type and size of the file after data processing, and how is these data used for visualization?
- What are the factors that affect the pollutant gas level, how to show these considerations in visualization?
- How to visualize the impact of the severity of the COVID-19 on the pollution gas content, and how is it being affected?

One hypothesis would be that as the pandemic spreads, each country experienced varying degrees of lockdown, therefore, the suspension of activity at industrial enterprises and transportation facilities, each pollutant level should show reduction.

These concerns and hypothesis were considered in the process of accomplishing this project, and the solutions and consequences will be shown in the next section. The conclusion section will include a summation.

4. PROJECT SETUP

In this project, we investigate the global impact of activity reductions resulting from the spread of COVID-19 on air pollution. For this purpose, we use the polluting gas data generated by the TROPOMI onboard the Sentinel-5 Precursor (Sentinel-5P) satellite, launched in October 2017. Settin up TROPOMI aims to collect the best quality data and collect data for a more extended period in air quality and climate research. The visualization of the TROPOMI dataset is conducive to meteorological scientists to observe climate and air quality changes to ensure that the best scientific data will influence climate research and policy formulation in the future [9].

4.1 Raw Data Investigation

4.1.1 TROPOMI Data

TROPOMI data set consists of the Level 2 products. The Level 2 products contain the gas of ozone, methane, formaldehyde, aerosol, carbon monoxide, nitrogen dioxide, and sulfur dioxide. The TROPOMI datasets are consist of Near-Real-Time stream (NRTI), Off-Line stream (OFFL), Reprocessed File stream (RPRO), and the Cloud Optimised GeoTIFF (COGT) files.

NRTI data stream is available within 3 hours of sense but may be incomplete or have quality defects. The OFFL data is generally available about a week after the NRTI data, and RPRO is the best quality version. In this project, we decided to observe the air quality changes between 2019 and 2020. However, the RPRO has much fewer data compared to the NRTI and OFFL files. Some gases file only contains the data in 2018 or 2019. NRTI, OFFL, and RPRO are in the netCDF file format. COGT files are in the TIFF image file format, generated from NRTI, OFFL, and RPRO data. COGT contains the image files, which have less other gas information. As a result, in this project, we decide to study and analyze the gas data of OFFL files.

OFFL folder contains NetCDF files for multiple gases from 2018 to 2021. The netCDF files are set up with group hierarchies, the METADATA group and the PRODUCT group, which store scanline (which indicates the dimensions of the satellite's flight direction), ground-pixel, time, corner, layer, longitude, latitude, quality value, and value array, etc. This project will only keep the most important data like the latitude, the longitude, the quality value, and the value array.

Through the list of S5P/TROPOMI Level 2 data products, data of CO, CH_4, Tropospheric NO_2 and SO_2 would be the domain of interest in this project. Data processing in the OFFL stream from 2019 to 2021 is focused on.

- Methane (CH_4). The file contains about 900,000 values on the column averaged dry air mixing ratio of CH_4.

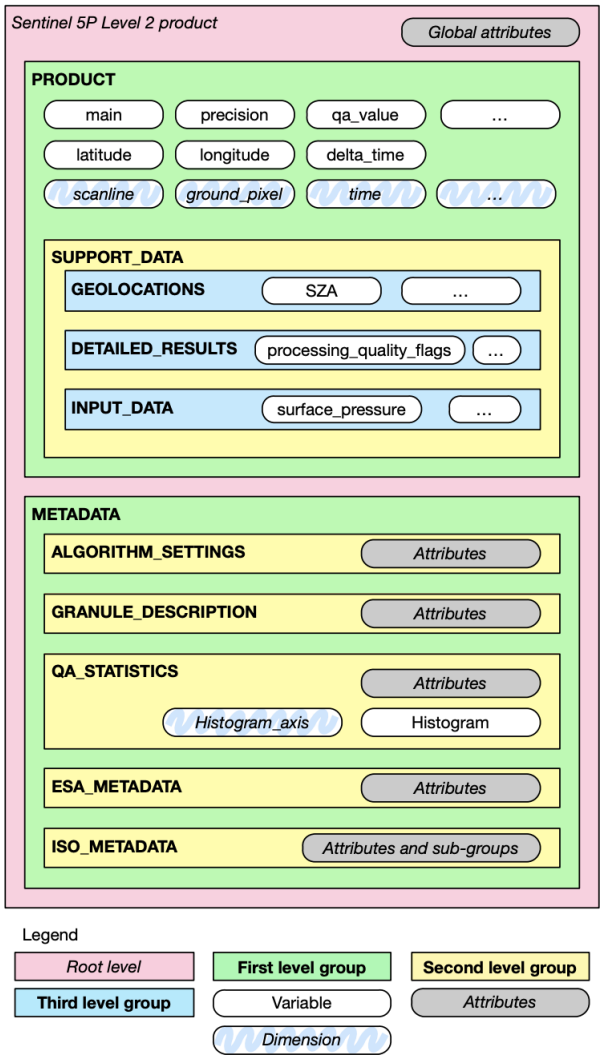

Figure 1: Graphical description of the generic structure of a Level 2 file.

- Sulphur dioxide (SO_2). There are approximately 1,80-0,000 values on the mole content of SO_2 in the atmosphere.

- Carbon monoxide (CO). The file also contains about 900,000 values, which is the mole content of CO in the atmosphere.

- Nitrogen dioxide (NO_2). There are 1,800,000 values per variable, the variables are the mole content of tropospheric vertical column of NO_2, the averaging kernel and troposphere air mass factor.

Figure 2 shows qa_value distribution in one piece of NetCDF file. As we can observe from the figure, each NetCDF file only represents part of the time of the day and part of the area of the whole world. If we want to read for complete one-day data, we must scan all the NC files in the same-day folder. As we mentioned previously, qa_value will influence the data quality. By deleting the substandard qa_value, the step will filter out most of the points. Figure 3 shows the

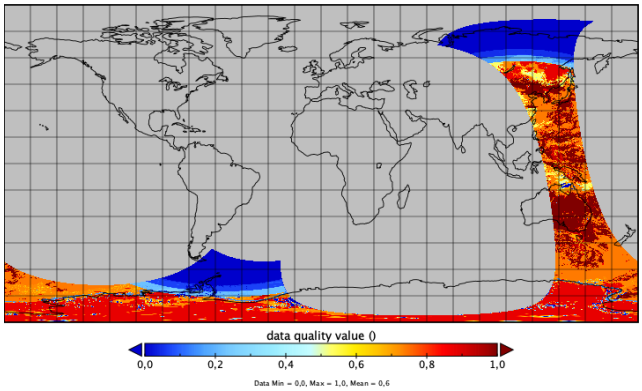

Figure 2: NO2 NetCDF-file qa_value distribution within range of (0,1) in 2019

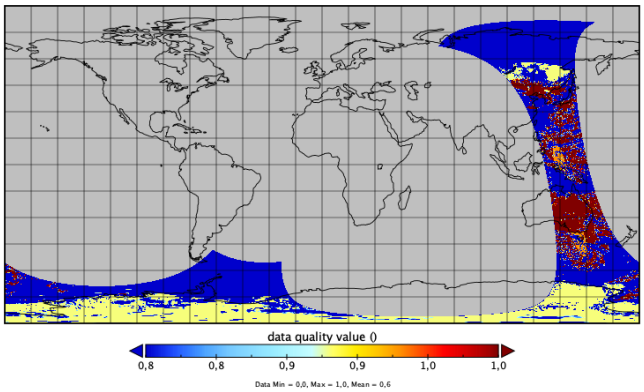

Figure 3: NO2 NetCDF-file qa_value distribution within range of (0.75,1) in 2019

qa_value distribution within the range of (0.75, 1). The result gives evidence that not every coordinate is meaningful and valuable to the project. We can also observe that only the country-level coordinates are detected. There also contains the point which is located on the ocean and place with no countries. Also, above the previous introduction, though the OFFL is more cleaning than the raw data set NRTI, there are still some meaningless and wrong values inside the dataset. All the above observations provide the idea and evidence to the data extraction section.

4.1.2 *COVID-19 Data*

The Oxford COVID-19 Government Response Tracker (OxCGRT) [14] collects information on several different common policy responses governments have taken, scores the stringency of such measures, and aggregates these scores into a common Stringency Index. The data set collecting is an ongoing project, which has been collected since January 21, 2021. The OxCGRT is a country-level data, it contains Entity (country name), Code (country code), Day (time), stringency_index(government response stringency index) four attributes. The OxCGRT data set is a CSV file format that can be easily read and analyzed by Spark. Below is the visualization of the Netherlands government response stringency index in 2020 and 2021. Line 2021 only has the data to October because the newest version of the data set is October 2021.

4.2 Data Process Pipeline

Based on sections 4.1.1 and 4.1.2, we will introduce how to extract necessary information from the raw data. The algorithms designed for this project should first test on the small dataset. After getting the program's correct and successful execution, the demo program can be applied in the databricks environment. In this way, we can guarantee both the correctness of the algorithms and the safe use of the data bricks. The algorithm mainly contains three steps. They are data reading and import, data extraction, writing to result to JSON and GeoJSON.

4.2.1 *Data reading and import*

There is a directory structure in the dataset. The system helps users and programmers to apply data discovery. This directory structure is based on the filename and is as follows: "XXXX/PPPPPPPPPP/YYYY/MM/DD." XXXX means processing stream (NRTI, OFFL, COGT), PPPPPPPPPP means Product identifier, YYYY means year, MM means month, DD means Day. For example, "OFFL/L2__NO2___/2019/10/01/". Each gas in the project will show as a month unit, so the first preparation before parallelization is to read all the file path names into a list. In the reading file section, the project in this paper chooses the os (Miscellaneous operating system interfaces) package. It is also necessary to pay attention that files don't end with ".nc." After appending all the NetCDF files' paths into the list, the next step will be parallelization.

4.2.2 *Data Extraction*

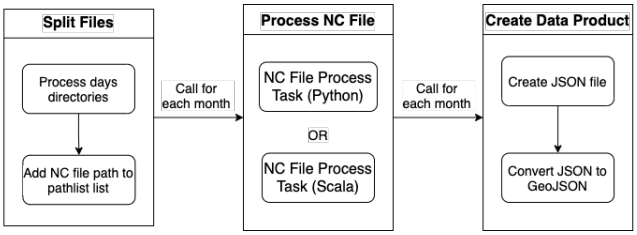

Figure 4: Data process pipeline.

In the project, two data extraction methods have been tested and compared. One of the methods was realized based on Python, and the other used Scala. After comparing the two methods' efficiency, the Python-based method is adopted in the project at last.

```
name| datatype| dimensions| shape
scanline | int32 | ('scanline',) | (372,)
ground_pixel | int32 | ('ground_pixel',) | (215,)
time | int32 | ('time',) | (1,)
corner | int32 | ('corner',) | (4,)
layer | float32 | ('layer',) | (50,)
delta_time | int32 | ('time', 'scanline') | (1, 372)
time_utc | <class 'netCDF4._netCDF4.VLType'>: string type | ('time', 'scanline') | (1, 372)
qa_value | uint8 | ('time', 'scanline', 'ground_pixel') | (1, 372, 215)
latitude | float32 | ('time', 'scanline', 'ground_pixel') | (1, 372, 215)
longitude | float32 | ('time', 'scanline', 'ground_pixel') | (1, 372, 215)
carbonmonoxide_total_column | float32 | ('time', 'scanline', 'ground_pixel') | (1, 372, 215)
carbonmonoxide_total_column_precision | float32 | ('time', 'scanline', 'ground_pixel') | (1, 372, 215)
```

Figure 5: variables details

As shown in the Figure 4, the first step is to select the variables. As mentioned in 2.1 and 2.3, most variables and

gases are not meaningful for the project. We will finally use the latitude, longitude, qa_value, and value array for data visualization. The project is using "netCDF4" package for reading NetCDF files. Figure 5 shows the detail of the variables of the PRODUCT group, including name, datatype, dimension, and shape. Each of the variable arrays will be transferred to a 2D array. The 2D arrays have the same data structure as each other. For the more profound explanation, each latitude in the 2D array corresponds to a longitude, a qa_value, and value. If one of the points is deleted in a 2D array, we will need to delete the other three corresponding 2D array points.

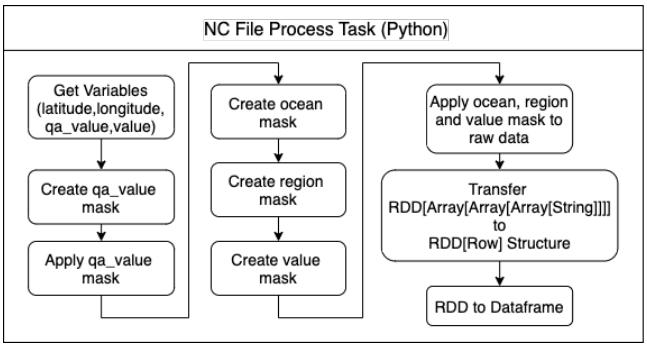

Figure 6: Process NC File(Python).

To realize the above idea, we used "NumPy.ma" to segment the target data. "NumPy.ma" is a tool package to apply the function on the masked array. A masked array is the combination of a standard "NumPy.ndarray" and a mask. The mask has the same data structure as the "NumPy.ndarray." The mask only contains two types of values. One is TRUE, and then another one is FALSE. If the mask's value is TRUE, it should remove the corresponding value on the ndarray. At first, we need to create a qa_value masked array. Which is mentioned in 2.3, NO_2's qa_value should be higher than 0.75. For the SO_2, CH_4 and CO, the qa_value should be higher than 0.5. The function "ma.maksed_less" and "ma.maksed_greater" will be used to create a masked array. Next, using the "ma.getmask" function can get the mask of the masked array. At last, we need to apply the qa_value mask to latitude, longitude, and value array. The mask applying step will remove all the cloud-cover scenes in the data set will be removed.

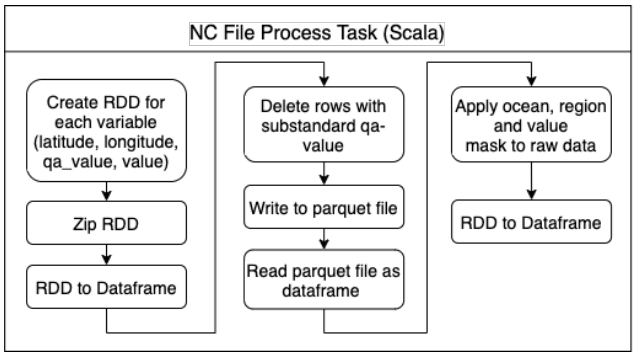

Figure 7: Process NC File(Scala).

Table 1: Data extraction result.

Folder Name	Original Size	GeoJSON Size	Reduce Rate
L2_NO2_2019	1900GB	2.53GB	99.86%
L2_NO2_2020	2200GB	2.51GB	99.88%
L2_NO2_2021	1900GB	1.72GB	99.99%
L2_SO2_2019	4000GB	1.91GB	99.99%
L2_SO2_2020	4600GB	1.89GB	99.99%
L2_SO2_2021	3800GB	1.43GB	99.96%
L2_CH4_2019	242GB	0.97GB	99.59%
L2_CH4_2020	284GB	1.03GB	99.63%
L2_CH4_2021	242GB	1.06GB	99.56%
L2_CO_2019	679GB	1.84GB	99.72%
L2_CO_2020	788GB	1.82GB	99.76%
L2_CO_2021	635GB	1.54GB	99.75%

The project plan and goal is to observe the air quality before, during, and after the COVID-19 among countries. Current data contains both oceans and dryland coordinates. It will be helpful only to extract the rows whose coordinates belong to the land. Creating an ocean mask will be a good idea to segment only the data in the land. The package "global_land_mask" will be imported to help judge whether the coordinates are on the ocean. Furthermore, some locations have abnormal values out of the range of standard latitude and longitude. So we also created a mask called location mask. When observing the data set, it is also noticed that some of the value of the value array is lower than 0. A negative number is a measurement or machine error. As a result, a value mask is also created. The ocean, location, and value masks will finally be combined into one mask called the final mask. The final mask will apply to each ndarray, which is already masked by the qa_value mask(Figure 6). We use the qa_value mask first and apply the final mask because the qa_value mask will delete 99% of the original data. The running time will also be reduced because of the smaller data set.

In the scala-based data extraction method, the "ucar.nc2. NetcdfFile" package is used to read NC files. There should be a corresponding RDD for each variable, latitude, longitude, qa_value, and value array. After creating 4 RDDs, they need to be zipped together using RDD basic conversion operation "zip." The next step will be to delete the substandard qa_value row. After cleaning the data, the dataset will write into the parquet file as a monthly unit. Next, we will reuse Python to read the parquet file and apply ocean, region, and value mask to the data. The new RDD will convert to Dataframe.

At last, we execute two data extraction methods. The executing time for the Python-based approach is for a maximum of 30 minutes, and a minimum of 1 minute, and normally for 10 minutes. The executing time for scala based method usually is slower than in Python. At last, we finally decided to use the Python-based method in this project.

After reading and filtering the data, the redundant variables in the netCDF file will be removed. The extraction process will reduce the amount of data for both computation and storage. Table 1 shows the data reduction result. As we can see, the reduced rate of data extraction is bigger than 99.6%.

4.2.3 COVID-19 Data Extraction

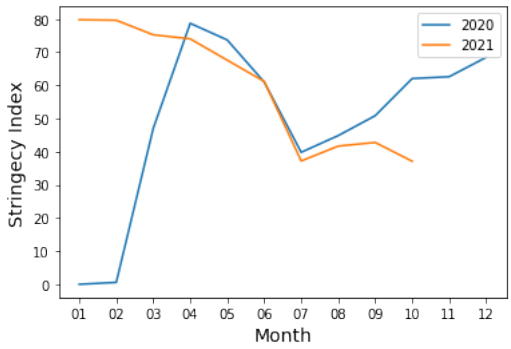

Figure 8: Stringecy index change in Netherlands

Three steps will conduct data extraction for OxCGRT data. Firstly, calculate the average stringency index by month unit. The final visualization will show the TROPOMI data set month by month, and the OxCGRT data should also transform into a monthly unit. The raw data of the OxCGRT is separate by day. We need to group the day for each month and calculate the average index. The second step should be to transfer the code attribute to coordinates. Each country code number is replaced by the country location value, which can be readable in the GeoJSON file. The detail will be explained in 4.2.3. Finally, the data result will also be written into a GeoJSON file.

4.2.4 *Creating GeoJSON Files*

To visualize data using the Mapbox Javascript API, the final input for the visualization product should be GeoJSON format. In the project, we used two GeoJson formats. One is for visualizing emission data, and the other is for visualizing the Government Response Stringency index.

The emission data GeoJSON file has the format as below. Each of the records in the GeoJSON file should contain information about latitude, longitude, and a value representing the heatmap color.

```
{
    "type": "FeatureCollection",
    "features": [
        "type": "Feature",
        "geometry" : {
            "type": "Point",
            "coordinates": [lon, lat]
        },
        "properties" : {
            "value": value
        }
    ]
}
```

The Government Response Stringency index GeoJSON file uses the format as below. It should be mentioned that, though the emission data is saved month by month, there is only one GeoJSON file for the Response Stringency index data. We use the properties "year" and "month" to mark which data should be in which layer. The value property in the GeoJSON represents the Government Response Stringency index.

```
{
    "type": "FeatureCollection",
    "features": [
        "type": "Feature",
        "geometry" : {
            "type": "Point",
            "coordinates": [lon, lat]
        },
        "properties" : {
            "value": value,
            "year" : year,
            "month" : month,
        }
    ]
}
```

4.3 Data Visualization

The final product of this project shows the changes in the levels of different pollutant gases measured by TROPOMI around the world from 2019 to 2021. It geographically visualizes the global impact of the COVID-19 epidemic on these emissions. This section describes how to create the data visualization website and the results of visualization.

4.3.1 *Creating Tilesets*

To effectively show the impact of COVID-19 on air quality on a global scale, the visualization product uses Mapbox to read GeoJSON files to create maps. Tilesets are the primary data format for Mapbox maps which is easy for caching and loads swiftly. Mapbox Tiling Service (MTS) allows users to create vector tilesets by providing source data with recipe transformation rules, which can set the "lowest_where_in_distance" to reduce feature density. As shown below, we use recipes to subdivide each tile at every zoom level into many equally spaced regions and only keep the average feature in each region. It helps us keep the most essential features at each zoom level while ensuring they are spatially distributed.

```
{
    "tiles": {
        "limit": [
            "lowest_where_in_distance",
            true,
            8192,
            "SCALERANK"
        ]
    }
}
```

Since the Mapbox style only allows a maximum of 15 sources, we use Multilayer tilesets to create a tileset with up to 20 layers. Each layer can have a unique tileset source, which can fit more data into a style. Therefore, Tilesets CLI was used to create tileset sources from GeoJSON files. We used Mapbox Studio Style to create a total of 251 layers containing monthly gas distribution layers, government lockdown levels layers and country boundary layers, all of which can be read and displayed in JavaScript and static HTML.

4.3.2 *Visualization Website*

The website shows a world map of the distribution of the NO_2, CO, CH_4 and SO_2 gases from 2019 to September 2021.

There are overall three time periods: the no outbreak period, the early outbreak period and the late outbreak period.

Figure 9 shows what the website looks like, each gas variable is positioned by its latitude and longitude to a pixel point on the map, and the colours of the pixel point depend on the concentration of the gas. Moreover, a time slider and two drop-down menus are set to allow the users to view the map of different gases at different times. In order to observe the effect of COVID-19 on air pollution, there are the options to display the OxCGRT of different countries and regions as in figure 10. The main causes of air pollution are from factory emissions and human activities, so government lockdown levels are more relevant to air quality than the confirmation rate of COVID. Stringency levels ranging from 0 to 100 can be displayed on the website as numbers and circles, and the size of the circle represents the level of lockdown, which can help observe the details of COVID-19's impact on air quality more effectively.

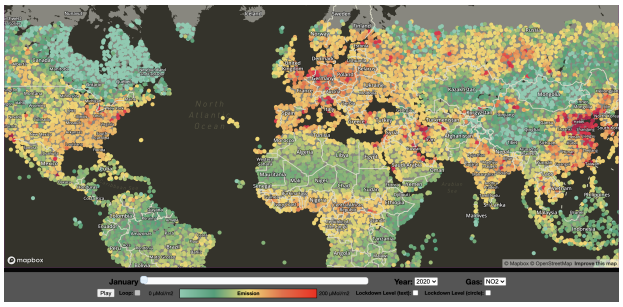

Figure 9: Website visualization.

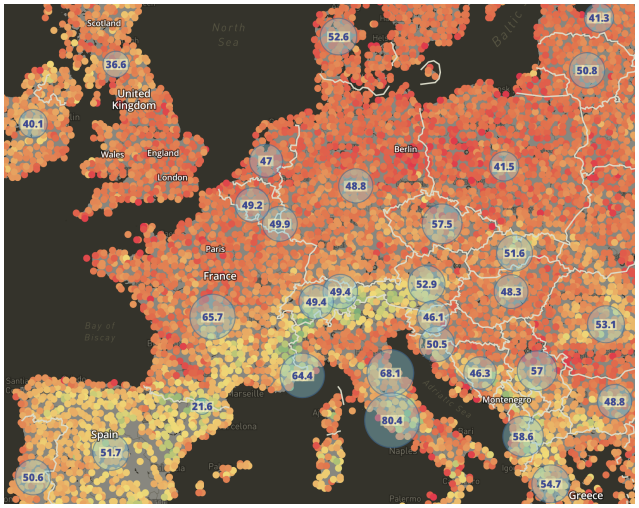

Figure 10: CO in March 2020 in Europe with lockdown level.

4.4 Result Analyzing

As pollution is mostly caused by major industrial countries, it is easier to monitor pollutant changes in those areas in relation to the lockdown. Afterwards, our group will concentrate on the locations with the greatest pollutant gas emissions. For example for NO_2 are these three regions: eastern Asia, Europe, and the east and west coasts of the United States. Since the season has such a significant impact on polluting gases, the statistics for 2020 and 2021 must be compared to 2019, which COVID-19 has not yet emerged in.

By comparative observation, the gas most affected by COVID is NO_2, and a rapid reduction in NO_2 emissions can be seen very clearly in the global response to the epidemic in 2020. Figure 11 shows the NO_2 in the atmosphere over the Netherlands in June between 2019 and 2021. Since the pandemic in China commenced in January, China has witnessed a bigger decline in NO_2 pollution in 2020 than in 2019, compared to other nations across the world. By February, the tendency had spread throughout Europe. Many European nations had a monthly average lockdown level of higher than 10.0. In March, the US's NO_2 emissions began to fall considerably, in combination with the steep increase in the lockdown level. Emissions in China, on the other hand, have climbed significantly since last month, despite the fact that the lockdown level remains near to 80. China's pollutants in April were similar to those in 2019, and the lockdown level was dropped as well. Emissions in Europe have gradually begun to rebound, and until September, they were not significantly different from the previous year which the lockdown level has declined from a high of nearly 90 to a low of less than 50. The United States has resumed NO_2 emissions since November. With the recurrence of the pandemic, the country has modified the lockdown level accordingly, reducing emissions for a period of time, such as in Europe in February and May 2021. However, emissions in 2021 are much greater in several other months than in 2019. One probable reason is that individuals are concerned of catching an infection when they take public transit, which leads to an increase in self-driving travels. Other possible factor that may contribute is the newly constructed plant.

As shown in Figures 13 and 14, Europe and China also show some reduction in CO concentrations in 2020 compared to 2019, and an increase in 2021 when the global response stringency is reduced. During the seasonal dry period in Africa, people ignite hundreds of fires to manage agriculture and grazing land for preparing fields for planting. A significant amount of CO was emitted throughout this process, negatively impacting local air quality [15]. However, through the comparison of the visualization results, The CO content has not been fixedly raised or lowered in comparison to the value in 2019, hence CO emissions in this region may not be influenced by the pandemic. Unlike other pollutants, the impact of CO on China can only be seen starting from March 2020 and recovered in May. One of the reasons may be that CO is mainly affected by transportation facilities. Although the usage of public transit has declined, this does not necessarily mean that the use of automobiles has reduced. Furthermore, the reason why the change of CO to other pollutants is not very drastic is due to its relatively longer life cycle in the atmosphere [7].

Figures 15 and 16 show the atmospheric concentrations of SO_2 during the cold season in Europe and China respectively, but from Figures 16 and 17 it can be seen that SO_2 emissions are mainly seasonal and latitudinal related. Heating begins in the north as the climate starts to cool, resulting in an increase in SO_2 emissions. The data of SO_2 cannot possibly be inadequate for direct comparison, unlike NO_2 comparison. Hence, prior to the European COVID-19 pandemic, SO_2 emissions in January 2020 were considerably

higher than in January 2019. This might be influenced by the weather or industrial development in that year. However, compare the differences between January and February of the two years, the reduction in SO_2 in 2020 is still evident, confirming the influence of the pandemic on the suspension of industrial factory activities. However, this drop is lower than that of NO_2 since the providing of heat remains a major source of SO_2 emission. With the improvement of the epidemic condition and the unblocking, SO_2 emissions steadily approached the value in 2019 in the next months. Furthermore, SO_2 outputs are less affected by the subsequent lockdown level changes.

However, not all gases were affected by the Corona virus and reduced. In Figures 18 and 19, Atmospheric concentrations of CH_4 increase rather than decrease in 2020 and 2021, Claus Zehner, manager from ESA's Copernicus Sentinel-5P, said: "One explanation for this could be that due to the reduced demand for the gas from COVID-19, it is being burned and emitted, leading to an increase in methane emissions in this area". On the other hand, it may be the gas that is least affected by human activity, therefore the COVID lockdown has little impact on its production. [6]

This data analysis is not comprehensive enough for considering just one year's dataset might be used as a comparison, which is somewhat uncertain. The fluctuations in pollutant gas levels will be more visible if the data is segmented on a weekly or daily basis. However, storing too much data will slow down the loading time of the web page.

5. CLOUD COST

Since all the processed data in the project will be stored in S3, there will be some costs during the project. According to Section 2.1, there are limited gases with obvious changes in the atmosphere affected by the outbreak, so our main subjects are the atmospheric distribution of NO_2, CO, CH_4 and SO_2.

For cost planning in cloud computing, the various call operations during data processing must be considered. For the S3 standard, PUT, COPY, POST and LIST will cost $0.005 per 1000 requests and GET, SELECT and all other requests will cost $0.0004. In our project, we need to fetch the files for the above gases in the OFFL path from the S3 server separately, then update the filtered and averaged new JSON files, and finally read the JSON files to write them into GeoJSON. The calculations show that there are 69,731 netCDF files for the selected gases in the OFFL folder from 2019 to 2021, during which time our team will be viewing this data multiple times, resulting in a cost of approximately $5 to access this data for the month. As we will eventually average the gas data on a monthly basis, the volume of processed data will be significantly less, with approximately 264 files being PUT.

The original file size of these gases from 2019 to 2021 is about 21270 GB, or 21 TB, after data filtering and data merging, the size of the files are expected to be controlled at 20.25 GB, which can effectively reduce the costs. We chose Mapbox as our visualisation tool and therefore did not incur any costs.

6. CONCLUSION AND DISCUSSION

The purpose of this project is to explore the impact of the spread of Corona virus on air quality. Therefore we selected data for four gases - NO_2, CO, CH_4 and SO_2 - from 2019 to September 2021, then extracted, combined, averaged and compressed the raw two-hourly sets of data. As an outcome, we were able to obtain monthly GeoJSON data files which were 99.5 percent less than the original file size. These files were then converted to tilesets, and Mapbox was used to visualize them. The timeline of the visualisation product is adjusted on a monthly basis, allowing users to observe changes in atmospheric content of various gases at different time. The severity of the COVID-19 is represented through lockdown levels.

The hypothesis in Section 3 is partially confirmed. As mentioned in 4.4, after comparative observations, NO_2, CO and SO_2 all showed a decrease in emissions during 2020, mainly due to the restrictions on human production and activities caused by the epidemic outbreak. While CH_4 appears to have the least relationship to the pandemic. By observing the changes in global air pollution, we can summarize some patterns and strategies. For example, the high SO_2 emission areas are mainly in the cold northern regions in winter, which shows the impact of burning fuel for heating on air quality, so we can use cleaner energy to control the emission of harmful gases.

We encountered some errors and problems during the project, and also found solutions with the help of our professor and the efforts of group members. In the data processing pipeline, as the netcdf file structure was quite specific and could not be directly converted to parquet file format, so we used two methods, the first was to use python RDD to read the chosen columns in the file and then convert it to dataframe format to filter data. The second method was to use Scala and zip the initially flat RDDs. The difference in runtime between the two methods is not very significant and depends mainly on the state of the cluster.

There is also a lot of room for improvement in our project. For example, the limited choice of gases, aerosols and O_3 are also well worth investigating. When processing the NetCDF files, we sometimes encountered some file corruption errors, which we resolved by skipping the file, but it would cause errors in the results. Besides, the three years of maps in the visualisation site need to be manually switched to compare, but it is rather confusing to put three maps on one page, so we can add pop-up windows to each area to show the change in data over three years, which can make the website more user friendly.

7. REFERENCES

[1] M. Bauwens, S. Compernolle, T. Stavrakou, J.-F. Müller, J. van Gent, H. Eskes, P. F. Levelt, J. P. Veefkind, J. Vlietinck, H. Yu, and C. Zehner. Impact of coronavirus outbreak on no2 pollution assessed using tropomi and omi observations. *Geophysical Research Letters*, 47(11), 2020.

[2] J. Braaten. Monitoring air quality with s5p tropomi data. https://medium.com/google-earth/monitoring-air-quality-with-s5p-tropomi-data-4f6b0aebe1c0.

[3] F. W. Cichowicz Robert, Wielgosiński Grzegorz. Dispersion of atmospheric air pollution in summer and winter season. *Environmental Monitoring and Assessment*, 189:1–605, 11 2017.

[4] ECMWF. Flawed estimates of the effects of lockdown measures on air quality derived from satellite observations. https://atmosphere.copernicus.eu/flawed-estimates-effects-lockdown-measures-air-quality-derived-satellite-observations?q=flawed-estimates-effects-lockdown-measures-air-quality-satellite-observations.

[5] A. Ernst and J. D. Zibrak. Carbon monoxide poisoning. *New England Journal of Medicine*, 339(22):1603–1608, 1998. PMID: 9828249.

[6] ESA. Detecting methane emissions during covid-19. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-5P/Detecting_methane_emissions_during_COVID-19.

[7] M. Filonchyk, V. Hurynovich, H. Yan, A. Gusev, and N. Shpilevskaya. Impact assessment of COVID-19 on variations of SO2, NO2, CO and AOD over east china. 20(7):1530–1540, 2020.

[8] K. H.J.Eskes. S5p mission performance centre nitrogen dioxide l2 no2 readme. https://sentinel.esa.int/documents/247904/3541451/Sentinel-5P-Nitrogen-Dioxide-Level-2-Product-Readme-File.

[9] K. N. M. Instituut. Tropomi. https://www.knmi.nl/kennis-en-datacentrum/uitleg/tropomi.

[10] Y. Ji, X. Deng, H. Liu, Q. Huang, K. Zhou, and Y. Tao. Temporal and spatial distribution characteristics of co total column over china based on tropomi measurements. In *2019 International Conference on Meteorology Observations (ICMO)*, pages 1–4, 2019.

[11] M. K. Kiln Farm. Nitrogen oxide (nox) pollution. https://www.britannica.com/science/air-pollution.

[12] S. Mahato, S. Pal, and K. G. Ghosh. Effect of lockdown amid covid-19 pandemic on air quality of the megacity delhi, india. *Science of The Total Environment*, 730:139086, 2020.

[13] J. E. Nichol, M. Bilal, M. A. Ali, and Z. Qiu. Air pollution scenario over china during covid-19. *Remote Sensing*, 12(13), 2020.

[14] A. Petherick, B. Kira, E. Cameron-Blake, H. Tatlow, L. Hallas, T. Hale, T. Phillips, Y. Zhang, S. Webster, J. Anania, L. Ellen, S. Majumdar, R. Goldszmidt, T. Boby, N. Angrist, M. Luciano, R. Nagesh, and A. Wood. Oxford covid-19 government response tracker. https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker.

[15] P. Przyborski. Carbon monoxide over africa. https://earthobservatory.nasa.gov/images/20471/carbon-monoxide-over-africa.

[16] L. Shikwambana, P. Mhangara, and N. Mbatha. Trend analysis and first time observations of sulphur dioxide and nitrogen dioxide in south africa using tropomi/sentinel-5 p data. *International Journal of Applied Earth Observation and Geoinformation*, 91:102130, 2020.

[17] SoCalGas. Sources of methane emissions. https://www.socalgas.com/stay-safe/methane-emissions/sources-of-methane-emissions?__cf_chl_jschl_tk__=ISS3UKpvW1MszCic9H-JuBYPGwqH1DEasKQR8WgsKU9k-1636129565-0-gaNycGzNCf0.

[18] N. Theys, I. De Smedt, H. Yu, T. Danckaert, J. van Gent, C. Hörmann, T. Wagner, P. Hedelt, H. Bauer, F. Romahn, M. Pedergnana, D. Loyola, and M. Van Roozendael. Sulfur dioxide retrievals from tropomi onboard sentinel-5 precursor: Algorithm theoretical basis. *Atmospheric Measurement Techniques Discussions*, 2016:1–79, 09 2016.

[19] T. Verhoelst, S. Compernolle, G. Pinardi, J. Lambert, H. Eskes, K. Eichmann, A. Fjæraa, J. Granville, S. Niemeijer, A. Cede, M. Tiefengraber, F. Hendrick, A. Pazmiño, A. Bais, A. Bazureau, K. Boersma, K. Bognar, A. Dehn, S. Donner, A. Elokhov, M. Gebetsberger, F. Goutail, M. Grutter De La Mora, A. Gruzdev, M. Gratsea, G. Hansen, H. Irie, N. Jepsen, Y. Kanaya, D. Karagkiozidis, R. Kivi, K. Kreher, P. Levelt, C. Liu, M. Müller, M. Navarro Comas, A. Piters, J. Pommereau, T. Portafaix, C. Prados-Roman, O. Puentedura, R. Querel, J. Remmers, A. Richter, J. Rimmer, C. Cárdenas, L. De Miguel, V. Sinyakov, W. Stremme, K. Strong, M. Van Roozendael, J. Veefkind, T. Wagner, F. Wittrock, M. Yela González, and C. Zehner. Ground-based validation of the copernicus sentinel-5p tropomi no2 measurements with the ndacc zsl-doas, max-doas and pandonia global networks. *Atmospheric Measurement Techniques*, 14(1):481–510, Jan. 2021.

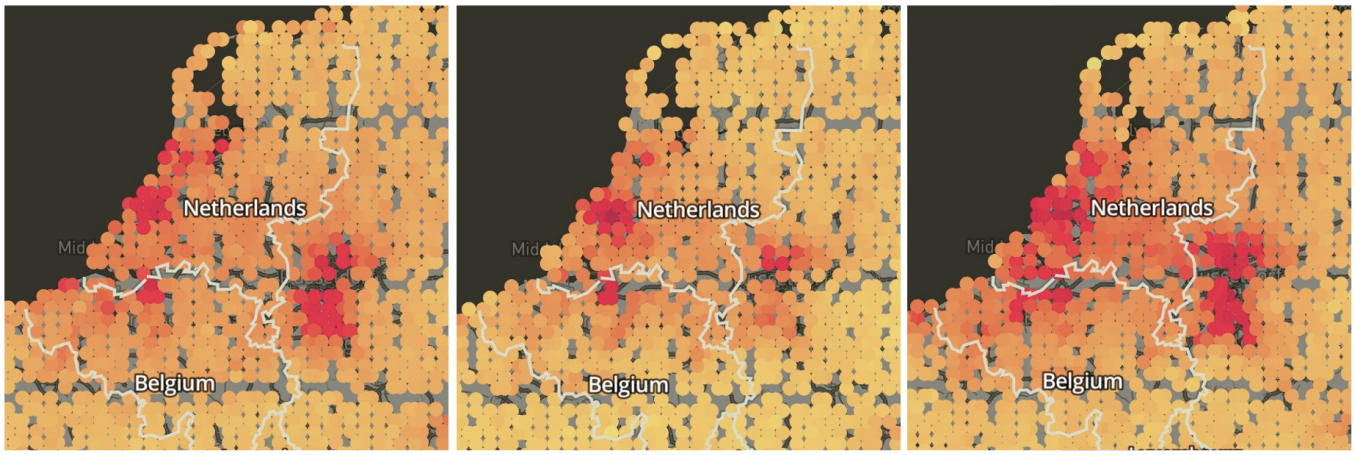

Figure 11: NO2 in June 2019 to 2021 in the Netherlands.

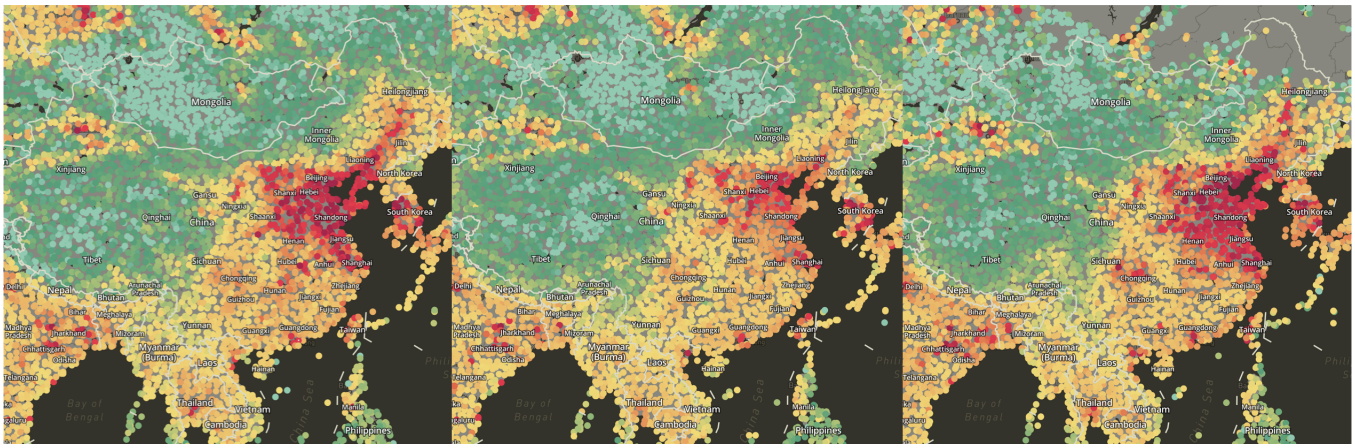

Figure 12: NO2 in February 2019 to 2021 in China.

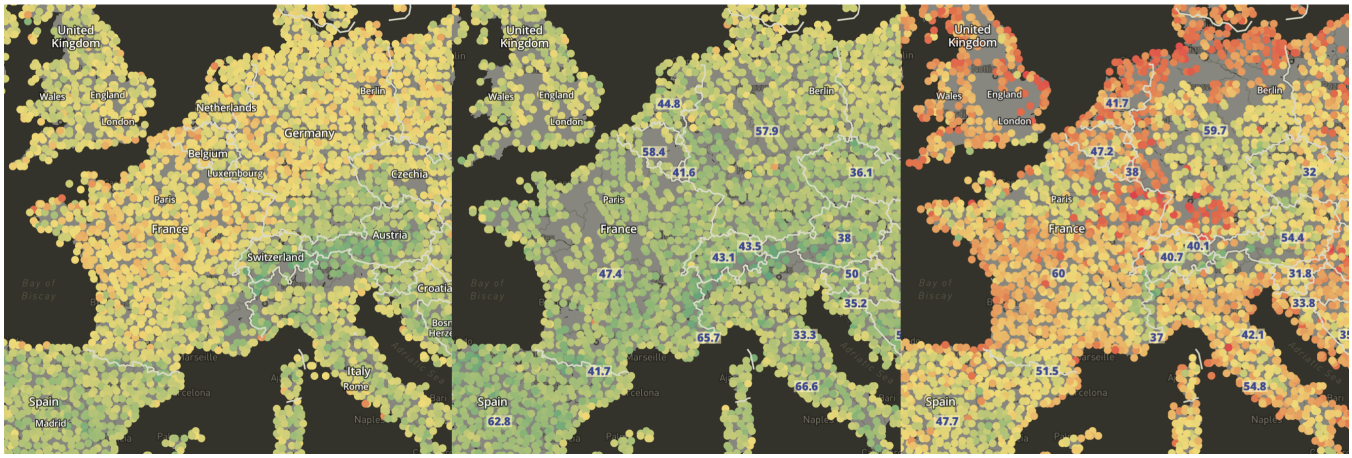

Figure 13: CO in August 2019 to 2021 in Europe.

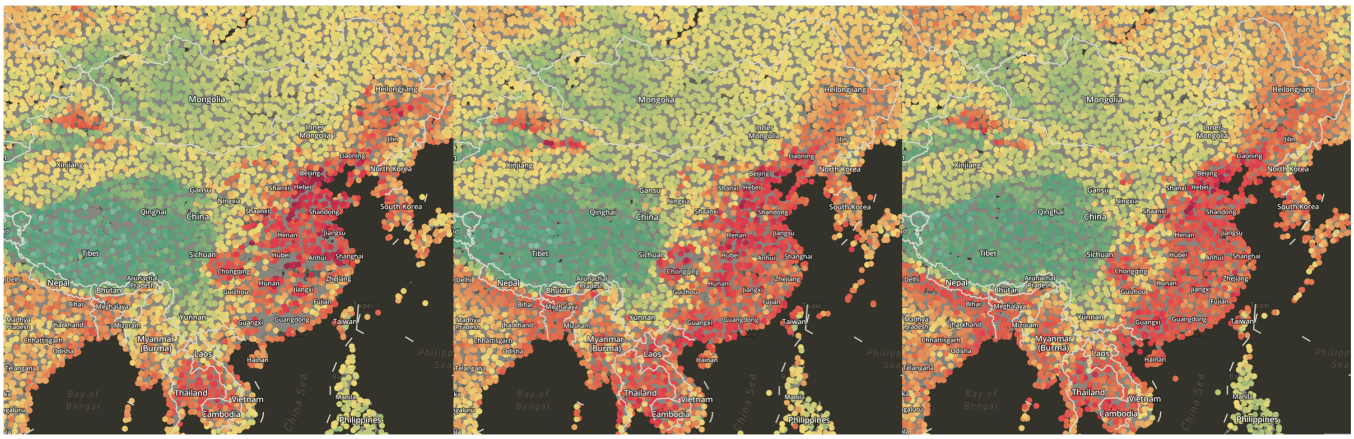

Figure 14: CO in February 2019 to 2021 in China.

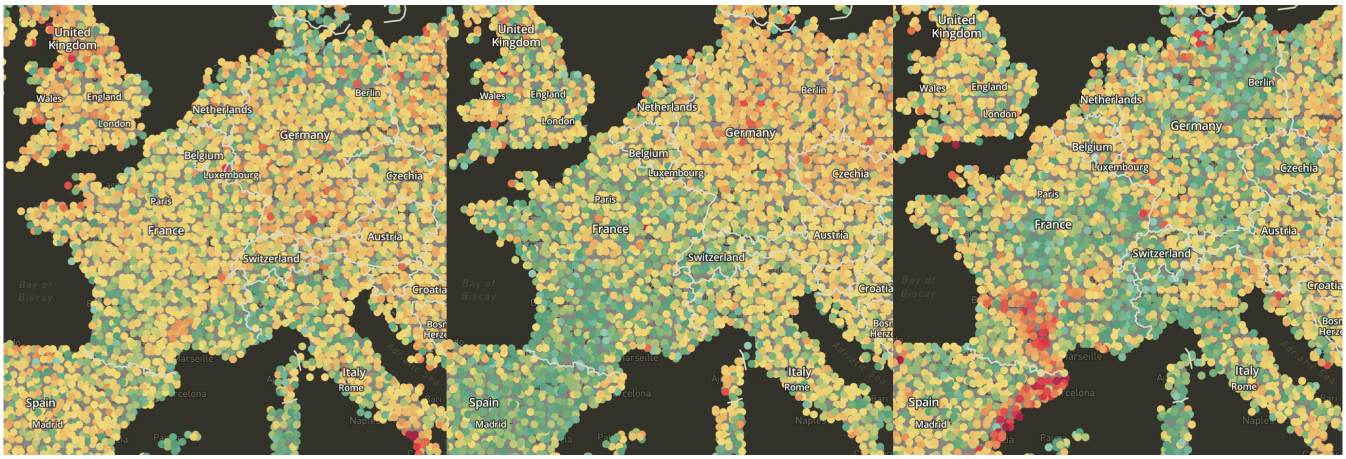

Figure 15: SO2 in September 2019 to 2021 in Europe.

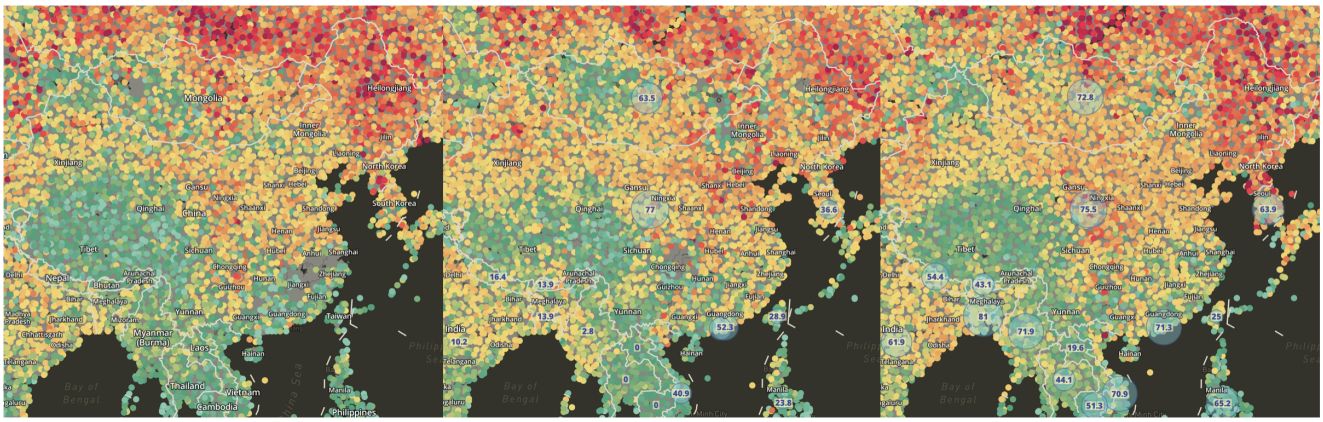

Figure 16: SO2 in February 2019 to 2021 in China.

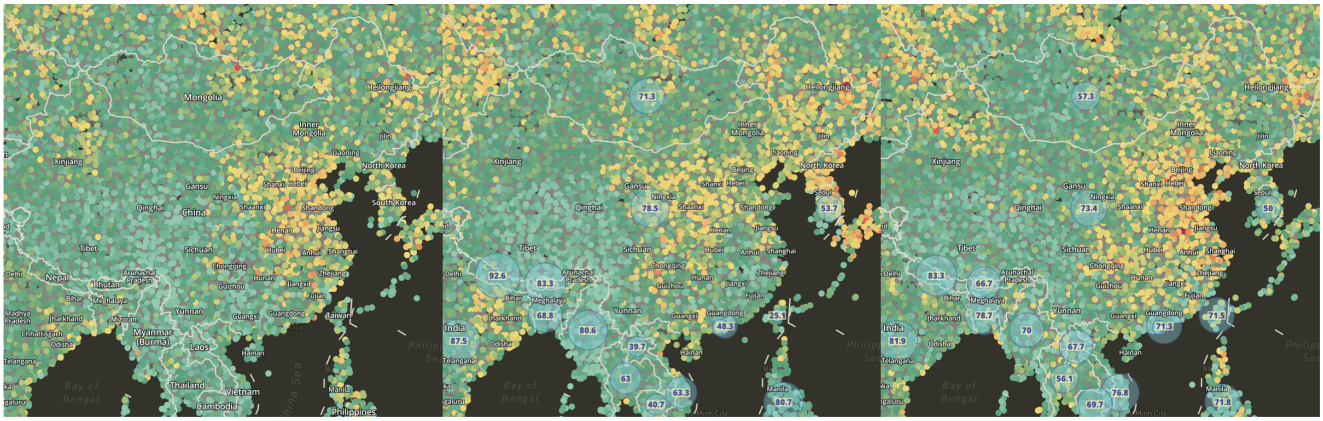

Figure 17: SO2 in June 2019 to 2021 in China.

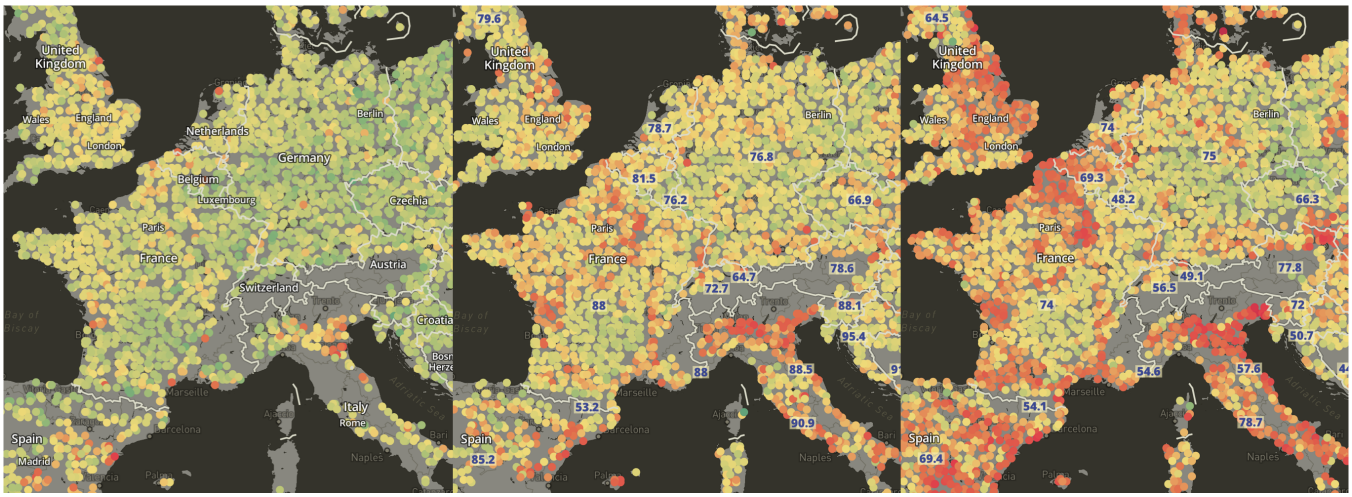

Figure 18: CH4 in April 2019 to 2021 in Europe.

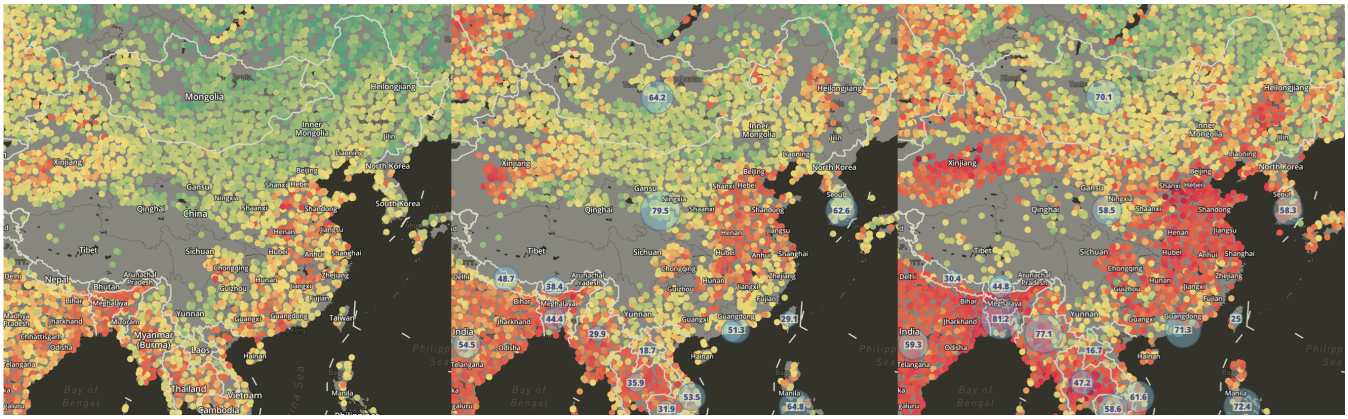

Figure 19: CH4 in March 2019 to 2021 in China.

Table 2: Project workload distribution.

Task	Who
Initial data investigation	All
Visualization tool search	Zhining Bai
NC file Visualization and Reading tool Search	Simei Li
Related work search	Yiran Li
OxCGRT related search	Yiran Li
Project pipeline design	All
Parallelizing the code on Databricks	Simei Li
Python code transfer to Scala	Simei Li
Data extraction	All
Data export	All
Processing OxCGRT dataset	Simei Li, Yiran Li
Writing JSON and GeoJSON files	All
Tileset creating method and tool	Zhining Bai
Creating Mapbox tilesets	Zhining Bai, Yiran Li
Adjust Visualization Styles	Zhining Bai
Visualization website implement	Zhining Bai, Simei Li
Report	All
Report formatting	All
Presentation	All