Interactive Data Visualization SS20

Project Group 32: Concept Paper

S. Aggarwal, S.Basu, A. Berneving, M.C.Dipinto

User and Task

The potential user of our visualization is an environmental analyst situated in Europe. Our visualization is meant to help the analyst with the task to assess if the Corona virus crisis has had any significant impact on the air quality for some of the bigger cities in Europe.

Datasets

For our visualization, we will use the following four datasets.

Air Pollution Dataset

Our main dataset is the COVID-19 Worldwide Air Quality data available from the Air Quality Open Data Platform. It contains air quality measurements of about 380 major cities in the world in the time period from January 2020 to today. Additionally, the same data is also available for the four quarters of 2019 (thus the entire year) and for the first half of the years 2015 - 2018, which allows us to compare the measured values before and during the crisis in comparison to previous years. For our visualization task, we will only consider the data for the cities located in Europe.

During each of the periods above, the data set includes average (median) daily measurements of the air pollution species PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3 (Ozone) and CO as well as the several meteorological variables (wind speed and gust speed, temperature, pressure, humidity, dew point, etc..) for each of the major cities. Additionally, each daily average is accompanied by the number of measurements that were used to calculate the median and variance as well as the minimum and maximum recorded value. One thing to note is that the reported air quality measurements are not the raw values. Instead, the raw concentrations have been converted into the US EPA standard

The data is given as a csv-file with point data (rows) containing nine data variables (columns). According to the Brodlie model, the dataset can be characterized as a E_9^P. The first four variables can be characterized as one ordinal (timestamp) and three nominal:

- *Date*: a UTC-based date string in the format YYYY-MM-DD.
- *Country*: the country in which the city is located, given as a two-letter code as defined in the ISO 3166-1 standard.
- *City*: the name of the city where the measurements were taken.
- *Specie*: specifies the type of data, as described above, that follows in the next columns.

Then, the following quantitative variables are given:

- *count*: number of individual measurements used to calculate the next variables.
- *min*: the minimum observed value.
- *max*: the maximum observed value.
- *median*: the median of the observed values.
- *variance*: the variance of the observed values.

The Air Quality Open Data Platform also provides a useful JSON-file containing additional information about the cities included in the data set. This will be used to extract the geographical location of each of the cities in the dataset.

Covid-19 Dataset

As our main source of the state of the crisis in each country we will use the collection of datasets made available by John Hopkins University on GitHub. It comprises three sets of time series data containing the number of confirmed cases, number of deaths and number of recovered for each country, respectively. By using this dataset, we are able to visualize the development of the corona virus in the country of each respective city. The user will be able to visualize the effects of Coronavirus crisis on the air quality parameters.

Each of the three datasets are available in the three csv-files

- time_series_covid19_confirmed_global.csv,
- time_series_covid19_recovered_global.csv, and
- time_series_covid19_deaths_global.csv,

and all follow the same characteristics. They are all built up as a mixture of point data and scalar arrays, where each record first contains the nominal data variables

- *Province/State*: The province or state (basically sub-region of the *Country* variable below) that the data belongs to and
- *Country/Region*: The country that the data belongs to,

and the quantitative variables

- *Lat*: Latitude of the country's location and
- *Long*: Longitude of the country's location.

Then, a scalar array containing a cumulative sum of the counts of each quantity (different one for each of the three files) follows. The counts are given in separate columns for each day, starting from 2020-01-22, and labeled using a date in MM/DD/YY format.

Using the Brodlie model, we can thus characterize the datasets as a $E_5^{\ P}$, where the last dimension itself is a $E_1^{\ nS}$ and n is the number of days since 2020-01-22.

Government Stringency Index Dataset

To additionally take the government's' actions into account in our visualization, for comparison with the air pollution metrics, we will make use of the government stringency index dataset available from ourworldindata. The index is a combined variable based on the actions taken by the government, such as school closures and cancellation of public events, and is given on a scale from 0 to 100, where 100 denotes the strictest.This will help us to relate to the Air quality index according to the strictness of the government. It is believed that stricter the government, the better is the air quality index during that time period in which the government restrictions are valid. The source of our dataset is from this visualization, and the data is available for all the countries but we are only interested in the data from the countries in Europe.

This dataset is obtained as a csv-file and contains time series data for each day starting from 1st Jan,2020 till June 2020. This gives us an added advantage that the user can observe how the stringency has changed according to the time, and thus whether it has any effect on the air quality or not?

The dataset contains 4 columns as different headers and has 28472 rows. According to the Brodlie model, the dataset can be characterized as a $E_4^{\ P}$. The columns can be classified as the following:

1. Nominal:
   Entity: Contains the name of the country that the data belongs to.
   Code: The code of the country as defined in the ISO 3166-1 standard.

2. Ordinal:
   Date: date in the format MMM DD,YYYY.

3. Quantitative:

Government Response Stringency Index: Given in the scale 0 to 100.

All the above columns are of interest to us. The datasets can be combined based on *Entity*, *Code* and *Date*, as most of the other other datasets have a timeline associated with them.

Mobility Dataset

To visualize possible changes in the habits of the residents of each country, we will also make use of the COVID-19 Community Mobility Reports made available by Google. These reports are based on anonymized location histories from users of Google's services that have been aggregated to provide a general idea of residents' behavioral changes during the crisis. The reason for including this dataset in our visualization is that it can be used to identify possible correlation between changes in resident's behavior and a change in the air quality. This could for example be caused by an increase in people working from home instead of driving to work every day.

Starting from 2020-02-15, the dataset contains relative changes from the baseline for the six categories *Grocery & pharmacy*, *Parks*, *Transit stations*, *Retail & recreation*, *Residential* and *Workplaces*, where the baseline is given by the median value, for the corresponding day of the week, during the 5-week period 2020-01-03 - 2020-02-06.

The dataset is provided as a csv-file with 13-dimensional point data, i.e., $E_{13}{}^{P}$ according to the Brodlie model. The main variables that are to interest for us are the following:

- *country_region_code*: Nominal variable for the two-letter representation of the country the data refers to, again given according to the ISO 3166-1 standard**.**
- *country_region*: the human name of the country the data refers to.
- *date*: Ordinal variable given as a date string in the format YYYY-MM-DD defining the date the data refers to.

Next, the quantitative variables regarding the actual data is given by

- *retail_and_recreation_percent_change_from_baseline*,
- *grocery_and_pharmacy_percent_change_from_baseline*,
- *parks_percent_change_from_baseline*,
- *transit_stations_percent_change_from_baseline*,
- *workplaces_percent_change_from_baseline*,
- *residential_percent_change_from_baseline*,

that contain the actual relative changes from the baseline as described above.

Visualization technique(s)

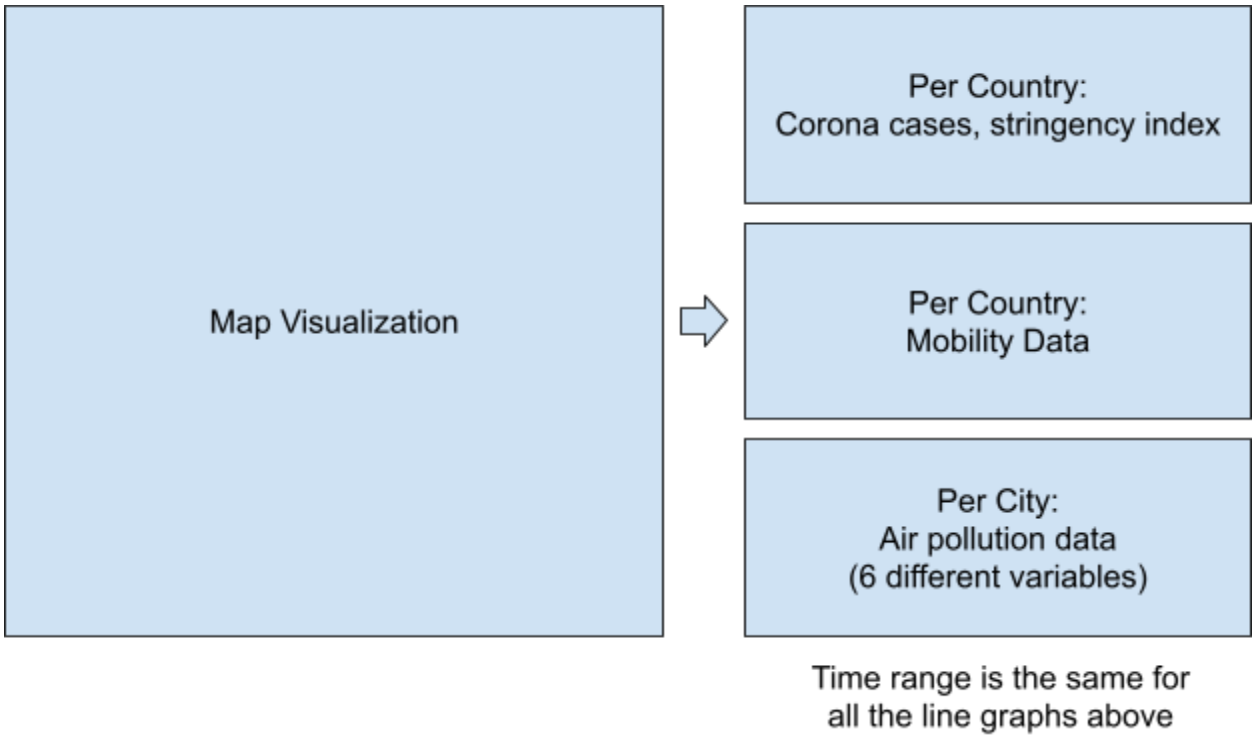
Our application focuses on showing the information which is relevant and useful for the environmental analyst. The main idea of our application is the effect of Covid situation on the environment. We are focussing on the European region only. The dashboard of our application shows the major cities of Europe on the map. The map is going to be interactive such that when the user clicks on any city/country, different charts and plots showing various parameters will be shown. The visualisations after clicking on a country are as follows:

- Visualisation 1: A line chart showing the trend in different Air Quality parameters (like PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3 (Ozone) and CO) of a city over time. The plot will be a time series chart. This visualisation will help the user in analysing how the Air Quality Index has changed during the covid times.

- Visualisation 2: We plot a time series line chart showing the government stringency index representing the restrictions put on by the government of the selected country, together with the number of active corona cases and deaths.

- Visualisation 3: It is a line plot showing the change in the mobility index of the citizens, for example the change in trends of mobility for public transport hubs. The chart will be shown as time series data.

When the user first clicks on a country map the 3 visualisation plots will be shown on the single screen. The user then also has an option to go back and select another country or a city to see visualisations related to that country.



Interaction

The interaction part of our project will give the possibility to the user to go through the visualization according what he/she is interested in: on the map he/she will be able to click on a specific european city and a line graph about Air Pollution, Government Mobility Stringency policy and Corona cases, of that selected city will pop up on the same screen of the map (one line graph for each variable).
The following actions will be implemented:

- On the map visualization:
    - mouse events (clicking) to see the data values of each city chosen
    - zooming in and out with specific button on the the screen to see data more clearly
    - panning

- On each line graph:
    - mouse events (clicking) to highlight specific coordinate variables

So the classes of interaction techniques that our project will include will be:

- *navigation*- we will use zooming, panning features on the screen space of the map to let the user control the visualization and scale or alter the view. In this way the user will be able to increase or reduce the screen space assigned to one or more focus areas.In our case the navigation is performed both on the *Screen Space* , because the panning,

zooming and rotating processes consist of pixel-level operation such as transformation or sampling, and on the *Data Value/Data Structure Space*, because zooming process involves focusing on particular data along the structure. The *Visualization Structure Space* is involved, too, because the user can explore the map by entirely moving through it.

- *selection*- we will let the user have controls for identifying a particular area of interest, isolating it on the map, to be the subject of some operations, such as highlighting details. In particular, by clicking on a specific city on the map the user will be able to see line graphs on the screen, one for each parameter chosen, such as air quality, mobility, status of covid in the country of that city,according to government stringency and active cases . In this way, the user will be able to know only what he/she needs in order to study the situation. In our case the selection moves on *Screen Space*, because it will be pixel-based according to the shape of the country which the user clicks on. But at the same time it moves on *Data Value Space* and *Attribute Space* in order to focus the attention on particular attributes of data, a particular selected subset of data, to be displayed, and on *Visualization Structure Space* because the user can select according the components (countries or cities) of the map visualization and according the graphs.

- *filtering*- By clicking on a specific city/country, the user will reduce the amount of data being mapped on the screen, to only show data relevant to that city/country. In this way he/she will move on the *Data Value Space* and *Attribute Space* in order to distinguish the variables he/she is interested in.