Data Visualization and Prediction of COVID-19 Global Cases

Samuel Svensson

Abstract— The COVID-19 (informally, 2019-nCoV) pandemic has become a global health emergency. The majority of the cases originate from China but as of writing the disease has spread world-wide and is quickly becoming uncontrollable. To understand in what ways this pandemic is spreading, a dashboard was implemented to visualize the geolocational data obtained from WHO and CCDC. This study aims to analyze what the most important data to be visualized is for a pandemic visualization dashboard, and what the most efficient way to present a large amount of data is.

Index Terms—Pandemic, Geovisualization, Geolocational and temporal data

1 Introduction

As COVID-19 continues to ravage the entire modern world it has become the center of attention from media and with it comes a feeling of uncertainty about the near future. It is difficult to grasp how serious it actually is and what kind of impact it will have. This report aims to present a geovisualization tool that provides a sufficient consciousness regarding the outbreak using current temporal data and visualization techniques. It has become imperative to evaluate a pandemic such as this because of its fast spread to adjoining countries. By developing a visual feedback to the common user, which is focused on geographical data, it is easier to interpret the data for what it is. With almost all media sources reporting on this on-going disease it is hard to understand how much it is affecting the world. There are lots of different variables that must be taken into account when visualizing a global pandemic such as this.

2 Background and Related Work

The visualization is partly inspired by the dashboard made by Johns Hopkins Center for Systems Science and Engineering (CSSE) at John Hopkins university. The main difference is possibility to move through time using the date slider to see day-specific data for a country. This was one of the main motivations when creating this implementation. Furthermore, there are several COVID-19 virus tracking tools available today but none seem to be as detailed and well-implemented as the Singaporean government dashboard [1]. It analyzes many different variables that are useful in seeing the whole picture and was an a contributing factor to why this dashboard was created.

3 Data

Taking into consideration that the pandemic is currently affecting much of the modern world, there are a number of different national departments and organizations that are collecting data. The obtained dataset was therefore a collaboration between the World health organization (WHO) [3] and Centers for Disease Control and Prevention (CDC) as well as the chinese CDC (CCDC) [4], to name a few. The collected data consists of three Comma-separated value (CSV) files with the most recent situational update report of confirmed, recovered and death cases around the world for each country. Some larger countries, such as China or USA, also had region-specific data to give a more detailed summary of COVID-19 cases. The dataset consists of temportal data with entries ranging from January 22nd of 2020 to present time. It also includes geographical coordinates (Latitude and Longitude) of each country/region together with the number of cases for that day in the area.

The CSV-files were converted into GEOJSON format which allowed for more flexibility during the implementation process. As mentioned above, China stands for a majority of the cases whilst other smaller countries only have a few dozen cases. This makes the data irregular when looking at it as a percentage of total cases worldwide. The only preprocessing needed for the data was the conversion to GEOJSON format since there were no null or undefined values. The three different datasets contained each around 450 rows (for each country and that country's regions) and around 60 columns (for each given date and positional data).

It is worth mentioning that the challenge of evaluating and visualizing the given data was not in the numerical values alone. It was rather a question of what kind of disinformation the dataset would present. The amount of confirmed cases a country presents might not accurately describe the actual number. More on this in chapter 8.

4 Method

To properly visualize the aqcuired temporal data a dashboard was created using Javascript and its subsequent libraries. The main functionality is the map, showing all the collected data, together with a corresponding date slider where the user can control which date the map will render data from over time. The map was created using Leaflet where circles of different radiuses were drawn onto the map. The radiuses are calculated depending on the amount of cases in each dataset. Since the amount of cases between countries were irregular (China being the largest contributor) [2] a logarithmic operation was done using the following equation:

$$2 \cdot \log_e r + 5 \tag{1}$$

where r is the calculated country-specific radius.

To accompany the Leaflet map a date slider was connected to it to make it possible to see date-specific information for a region. Three buttons were implemented so that the user could switch between the datasets of confirmed, recovered and cases of deaths. The map was positioned in the middle to further emphasize that the geovisualization was the most vital information in the implementation. The user can also zoom in on various continents and hover above the circles to see the name of the region together with the corresponding amount of cases.

A date slider which is positioned above the map controls which day of the chosen dataset will be shown. It is possible to drag the slider and see the map get updated instantenously. This is a very helpful functionality since it allows the user to see how the spread has developed over time.

To the sides of the map three lists where rendered to show the total amount of cases in each country. Countries which had region-specific data were summed and displayed as one single entry in the list. Each of the entries has a colored background with low opacity which fills up with an amount depending on the percentage of the total cases the country has, similarly to a horizontal bar chart. The lists are also connected to the map in a such a way that whenever the user can click a certain country entry. The map will then zoom in and center on the clicked country so the user can easily see how it compares to adjoining countries.

A line chart showing another perspective of how quickly the global cases had grown was implemented. Each line graph represented each dataset together with two different predicitions, a linear and an exponential prediction showing what the amount of confirmed cases could look like in 5 days from the latest update. The predictions follow a linear least-squares fitting method for the user to compare with the actual data.

5 Implementation

When implementing the dashboard Javascript was used together with the following libraries D3.JS, Chart.JS, Leaflet and Regression.JS. All development was done using the code editor Visual Studio Code with a Python HTTP server as a local server.

5.1 Date slider and map

The date slider was implemented using D3.JS which features integration for Leaflet maps. Everytime the user drags the slider the map updates the circle radiuses in each location accordingly. There are three buttons from which the user can select the different datasets to trigger a complete redraw of the map with the selected data. Upon clicking the button, the button color, circles and date slider change to different colours to be able to distiguish one dataset element from another. This is one form of strong visual feedback that the dashboard provides the user with.

5.2 Charts

To provide the user with an alternative form of data visualization, a chart was implemented using the library Chart.JS. Because the dataset was comprised of temporal data, a simplistic line chart was chosen to represent the data over time. There are five different plots showing the confirmed cases in bright yellow (neutral color), recovered cases (green, positive color), deaths (red color) as well as two prediction line charts. The prediction colors were chosen to be of the same nuance but still visibly different so that the user can distinguish them from eachother. These charts are also dotted instead of being solid. The difference is further clarified in the legend of the chart where it is easy to distinguish what data each line chart corresponds to. By clicking on a specific legend title it is possible to disable one or more line charts to isolate data or compare selected charts. The user can also hover on each datapoint to see the number of cases together with the date. It is also possible to switch the viewing mode to use logarithmic scale on the y-axis for a different perspective.

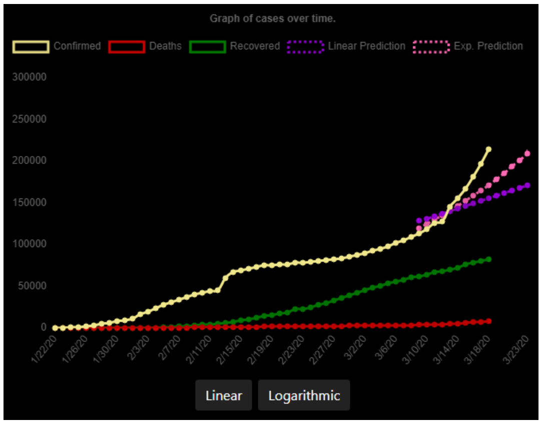


Fig. 1. Line charts with total amount of cases as well as future predictions

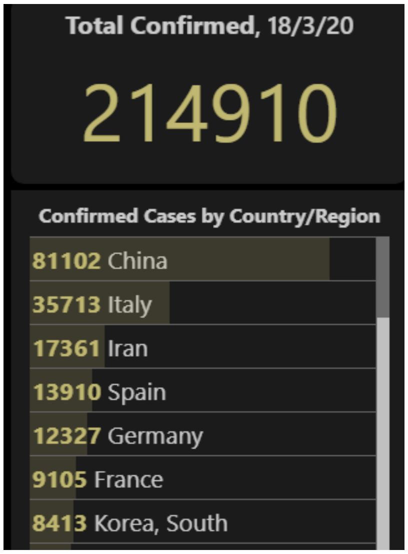
5.2.1 Data mining

As a data mining effort, the Javascript library Regression.JS was used to create a linear least-squares fitting method for visualizing both a linear prediction and an exponential prediction. The curve fitting method works by the inputting all numerical data from which a regression following the line equation is returned. The equation in the form of $y = mx + c$, where m and c are the coefficients. The same approach is used when calculating an exponential curve, but with the equation $y = a \cdot e^{bx}$ where a and b are the coefficients. This type of method is helpful for graphically showing how the data points are related to one another whether it is in linear or non-linear model. The method is based on R-Squared which is a common statistical measure used in regression models.

The model will then return all the new values which follow these equations and two new line charts can be drawn. To be able to do a prediction a few days ahead, some extra values were needed to be calculated and added into the array before plotting the line charts.

5.3 Numerical lists of total cases

As mentioned in chapter 3, some countries had more detailed, region-specific geolocational data available. As of writing, these countries are China, USA and Canada. When implementing the three lists with total number of cases, we summed up the amounts in each of these regions of those countries to present a better general overview for the user. Above each list the global total amount of confirmed/recovered/death cases is shown. Each country in the list is clickable which will trigger the map to zoom in and center on the selected country. As an extra visual feedback, each country also has a slightly tinted background which is partly filled up, depending on the percentage of cases that country has with regards to the total amount. These lists also follow the above mentioned color palette to make it easier for the user to separate the data visually.



Total Confirmed, 18/3/20

214910

Confirmed Cases by Country/Region

81102 China
35713 Italy
17361 Iran
13910 Spain
12327 Germany
9105 France
8413 Korea, South

Fig. 2. List of confirmed cases along with global total

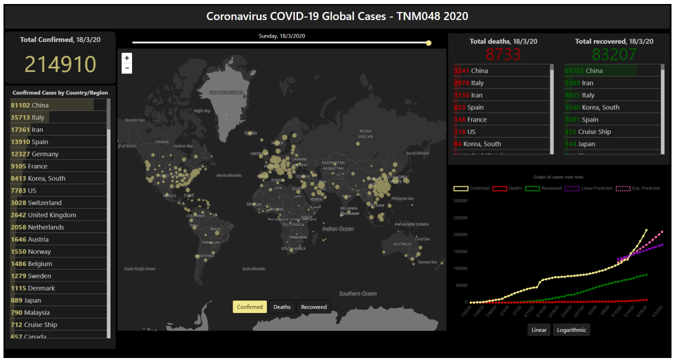
6 Results

A COVID-19 visualization dashboard was designed and implemented using Javascript and several libraries. These work to enhance the user experience and display all of the most vital data in a clear and concise way. The dashboard consists of a map with the geolocational data displayed as circle markers with different radiuses depending on amount of cases. Three buttons allows the user to switch between datasets and the appearance of the dashboard changes accordingly to familiarize the user with the chosen colors. A date slider above the map can be used to show temporal data for a specific day. It is also possible to drag it and see the data get updated continously. This can be particularly helpful when trying to understand how quickly the disease spread from China to the rest of the world.

Beside the map and slider are three lists containing the global amount of cases for either confirmed/recovered/death cases together with country-specific amounts. The countries are clickable which navigates the user to the selected country location where it can use the slider to see how the amount of cases grow over time.

A chart containing five line charts was implemented to give the user an alternative way of looking at the data. One line chart for each dataset as well as two predictions are visible in colors corresponding to the rest of the user interface. The charts can be disabled temporarily by clicking on the legend. This helps the user to isolate data and compare line charts with eachother. It is also possible to view the line charts with logarithmic axis scales.

The final dashboard result can be seen in Figure 3 below.



Coronavirus COVID-19 Global Cases - TNM048 2020

Total Confirmed, 18/3/20
214910

Confirmed Cases by Country/Region
81102 China
35713 Italy
17361 Iran
13910 Spain
12327 Germany
9105 France
8413 Korea, South
7783 US
3028 Switzerland
2642 United Kingdom
2058 Netherlands
1646 Austria
1550 Norway
1486 Belgium
1279 Sweden
1115 Denmark
889 Japan
790 Malaysia
712 Cruise Ship
657 Canada

Sunday, 18/3/2020

Total deaths, 18/3/20
8733
3241 China
2978 Italy
1135 Iran
623 Spain
148 France
118 US
84 Korea, South

Total recovered, 18/3/20
83207
69755 China
5389 Iran
4025 Italy
1540 Korea, South
1081 Spain
325 Cruise Ship
144 Japan

Fig. 3. Final COVID-19 dashboard to display a general overview of cases around the world

Because of the short time span for the project there were some limitations which include the amounts of cases in the sidebar lists not updating when dragging the date slider. This would have been an appreciated functionality but there are some downsides with it which was the reason for it not being prioritized in the implementation. The main negative side of it would be that the live updating of the data would interfere with the overall user experience. It could become too distracting from the main intention of the dashboard - to give a general overview of the current disease situation. This will be furthered discussed in chapter 8.

7 Evaluation

Structured interviews were conducted with 6 participants to evaluate the current implementation. This evaluation was to retrieve feedback of the current state of the dashboard and hear about possible shortages in the implementation. Furthermore, the interview was meant to see how efficiently the user understood how to interact with the tools and how intuitive they felt.

Every participant did the same procedure to ensure that anyones experience did not differ from the others. The evaluation begun by giving the user 2 minutes to freely interact with the dashboard, without any further instructions. This was to let the user acquaint themselves with the software and try to figure out how to see different parts of the data. Then several questions were asked and the user was also given tasks to perform. One for example was to pick a date, find a location on the west coast of the USA and see how many deaths there have been there as of today. Other tasks were more specific and gave less options to the user such as: "Find the number of confirmed cases Norway had on february 20th". This tested the users ability to navigate the interface and tools at hand.

After the user had performed the given tasks they were asked some questions in the form of an interview. The questions ranged from asking the user what the overall impression was of the dashboard, to being asked to give concrete examples of functionalities that may be improved. One of the questions was if the info was at any time unclear, which was an important question since it was tied to the aim of this entire study.

The results of the evaluation were overall positive. Some users wanted the data to be more integrated with the date slider, so that all of the country-speciifc numbers in the lists would update accordingly when dragging. One user would have wanted to get more country-specific data when clicking a country in the list. Also the lists could have contained regions and a search bar to look a particular country. A majority found the user interface very clean and the choice of colors was good as it did not strain the eyes. When asking the user if they thought it was too much information or if it was too cluttered it was not a problem. This was a clear indicator that the aim of the project was achieved.

8 Conclusions and Future Work

In conclusion, the resulting dashboard was appreciated by participants of the evaluation and several ways of visualizing and interaction tools were implemented to give a clearer overview of the current pandemic. The design and user interface gives the user an unobstructed view of what kind of data is the most vital when trying to grasp the situation as of today. There are of course areas that could need more work and maybe a different approach. Due to the time span of the project this was not possible but there are several ways the dashboard could be improven.

As mentioned in 3, there are some discrepancies in the data. The amount of confirmed cases might not be entirely true, since countries don't have the same testing capabilities. Sweden for example, recently decided to stop testing sick people unless they were hospitalized. Therefore, it might be more reliable to instead look at the deaths and try to predict amount of confirmed cases from that data. Perhaps a correlation between amount of deaths, country population and the current mortality rate of the disease could be used to create a different prediction of how many persons are actively infected.

Furthermore, the data only contained the number of different cases. There was not any detailed data on other properties such as gender and age distribution, transmission data (infection source), and infection length. It is known that countries all have different age distributions in their population, for example Italy. Therefore a conclusion could be drawn that the mortality rate would be higher there, especially if the disease appears in areas with a high amount of older people. This could have somehow been emphasized when clicking a country to see specific data.

The map is the main functionality and does an excellent job of displaying the geolocational data and the slider is a good way of viewing the data over time. Something that would have been interesting would be the ability to see a cloropleth map with how many new cases a country has received during that specific day, where countries with a low amount of cases would have a green color. The countries which has been hit with high numbers of new cases would be tinted from yellow if mild, to red as being severe.

Regarding the chart, an option would be to isolate cases in mainland China to the rest of the world to be able to compare. Also there could have been more detailed information on the tooltip when hovering a point in the chart, such as the percentage of increase of cases compared to the day before.

There are as mentioned several things that are open for future work and improvement for the dashboard. As this study has clearly demonstrated is how important visualizing pandemics has become to give people a better understanding of how fast it can become dangerous.

References

[1] *Tracking the Epidemic*, CCDC
https://co.vid19.sg/

[2] *Coronavirus disease (COVID-2019) situation reports*, WHO
https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[3] *Tracking the Epidemic*, CCDC
http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm

[4] Samrat K. Dey, Md. Mahbubur Rahman *Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach*, Department of Computer Science and Engineering, Dhaka International University , Acquired 2020-03-15
https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.25743