530 Final Project Report

Jordan Buckmaster, William Reed
buckmast@purdue.edu, reed226@purdue.edu

May 3, 2020

1 Datasets

For the COVID-19 dataset, we used the time series data for infections, recoveries, and deaths, provided here (from the repository mentioned in our project proposal):

https://github.com/CSSEGISandData/COVID-19

Additionally, we visualize population density to observe how it corresponds with COVID-19. For this data set, we use data from the Socioeconomic Data and Applications Center (SEDAC):

https://sedac.ciesin.columbia.edu

Typical migration patterns may also play a role in country to country transmission of COVID-19, so for this data we use information provided by the United Nations:

https://www.un.org

Since there is speculation that COVID-19 might become a seasonal infection, we are also interested in trends between COVID-19 and regional climate to see if it has a tendency to spread more in certain conditions, using data provided by WorldClim:

https://www.worldclim.org/data/monthlywth.html

2 Motivation

The question we hope to answer is with our project is: What are the largest risk factors contributing to the spread of COVID-19? COVID-19 is a very relevant issue in the world right now, and creating a visualization to understand the exponential nature of its spread and the contributing factors can be helpful in

fully understanding the virus and what actions are required to stop it. We believe that understanding the virus with visualizations can go a long way in mitigating panic and allowing people to think rationally about the situation. We also want to plot regional climate data to identify if COVID-19 spreads more frequently in certain climates. This will hopefully give us insights on whether COVID-19 could potentially be a seasonal virus or not. Aside from climate data, we will also visualize migration data and population density, two factors that we believe largely impact the speed of the spread of COVID-19.

3 Approach

There are various different components we definitely wanted to implement in the project, specifically infection spread, a density map, a climate map, and a migration patterns visualization. We chose to split these components among the two of us and develop them separately in separate python scripts. Once we had implemented all of the components separately, we combined them into one combined visualization script that can be used to visualize everything at once. We also originally planned to add on option to view the visualization on a globe, but decided against this in the end because we did not think the globe would add much to the visualization.

To visualize the infection spread, we opted to use a date slider to inspect the state of the spread at various points in time. The data we used included the longitude and latitude of the various data points, so we converted these to a 2D coordinate system and plotted them on a 2D map of the world. The size of the circle at each point on the map indicates the amount of infections. Because some data points (the US for example) have exponentially more cases than other data points, we used a logarithmic scaling on the size of the circles so that the smaller data points were not too small. We also chose to allow filtering out the infections, recoveries, or deaths data points so we could inspect each category more closely if desired. The legend is updated as the date is changed, as the number of cases corresponding to a size of a bubble changes as the date changes.

To visualize the population density, we chose to use a heatmap with a color scale that shouldn't interfere too much with the COVID-19 data. That is, lower values use colors in the blue-yellow range, so as to keep them distinguishable for the red and green data points used to show COVID-19 infections and recovers. Only the higher values in the density map utilize red, and due to the nature of population density, these densely populated areas do not show up too strongly on the map, keeping the COVID-19 data easy to see. The scale is logarithmic so as to make it easier to see a range of population density for lower populated areas, otherwise the scale is dominated by densely populated regions.

To visualize climate data, we chose to use several heatmaps, with one for each month's average minimum and maximum temperatures for the world. The

heatmap displayed is tied to the same time slider as the one used for COVID-19 data, so that when the month changes the climate heatmap can be updated appropriately. To reduce visual clutter with the heatmaps, only one can be selected at a time. The provided dataset was also too large to quickly load and display when the time slider changes month, so we also had to manually reduce its granularity to have it load faster. We also reduced the size of the density heatmap to match.

To visualize migration patterns, lines were drawn between countries where one country has immigration from the other. No direction was given on the lines because it is difficult to tell direction anyways with so many lines being drawn per country. With this mind, the purpose of the migration data is to visualize a general association between countries. Opacity indicates the amount of migration, so countries with a lot of migration between them, like the United States and Mexico, have more opaque lines. This also means that countries with immigration in both directions will have a significantly more opaque line due to two lines being drawn on top of each other compared to countries that only have migration in one direction. Countries with relatively little immigration between them (5% of the countries with the most immigration between them) do not have the corresponding line drawn to reduce visual clutter. Also, the provided data set didn't give longitude and latitude for each country, so we used our own translation CSv to map the names to longitude and latitude.

4 Problems

There were a few areas aspects of our visualizations that could be improved if we had more time to develop. Since there is no ability to filter out data points above/below a certain value, it can be difficult to inspect specific data point in densely packed sections of the map such as Europe. This can be mitigated simply by adding a slider for the minimum value of a data point and the maximum value. Also, the size scaling on the infection data points could be improved; as things are now, there is not much variation in size of the circles for the largest data points. For example, despite the recoveries and deaths in the United States being an order of magnitude smaller than the infections, the circle size does not appear significantly smaller on the visualization. This is simply due to the nature of logarithmic scaling, but we believe there could be a better way to visualize this.

It would also be beneficial to get a more recent data set for migration data. Currently, the migration and climate data is taken from an older data set from before 2020. With climate data, we can mostly extrapolate it to 2020 and not expect a big difference in the results, but it could be useful to get more fine-grained migration data for 2020. With our current visualization, we are only capturing general migration patterns that do not change with time, and while

this is good for determining countries that generally have a lot of migration between them, being able to see fine-grained day-by-day travel data for 2020 could be a bit more insightful.

The data that we found on immigration also only included a handful of countries. Countries where the immigration data that could be important for visualizing its affect on COVID-19, such as China, did not have data available, so we could not visualize immigration going in to the countries (emigration, however, could be captured by other countries' migration data). Ideally, we would use a more complete immigration data set, however such a data set is not available since some countries don't report their immigration.

5 Results

The following screenshots show key correlations that we try to capture between the early spread of COVID-19 and the risk factors that we identified:

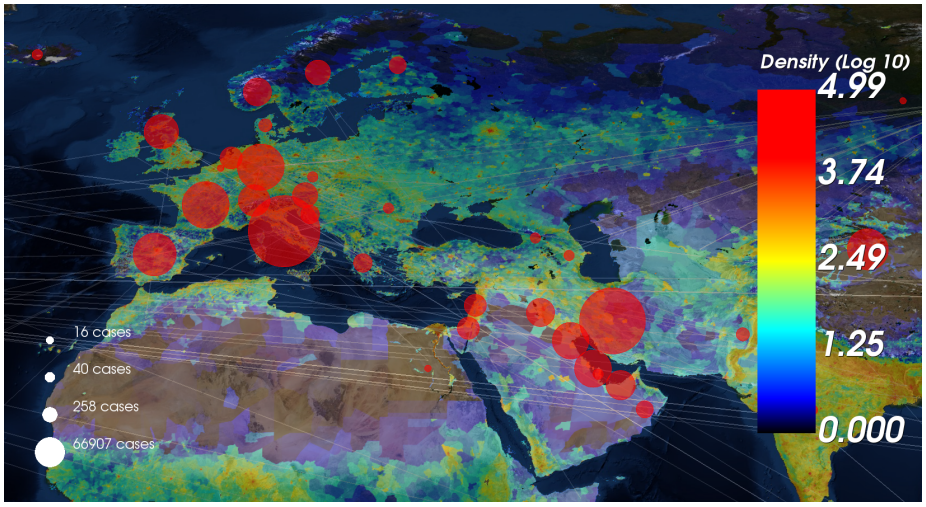

Figure 1: March 1st: Spread of COVID-19 in densely populated countries

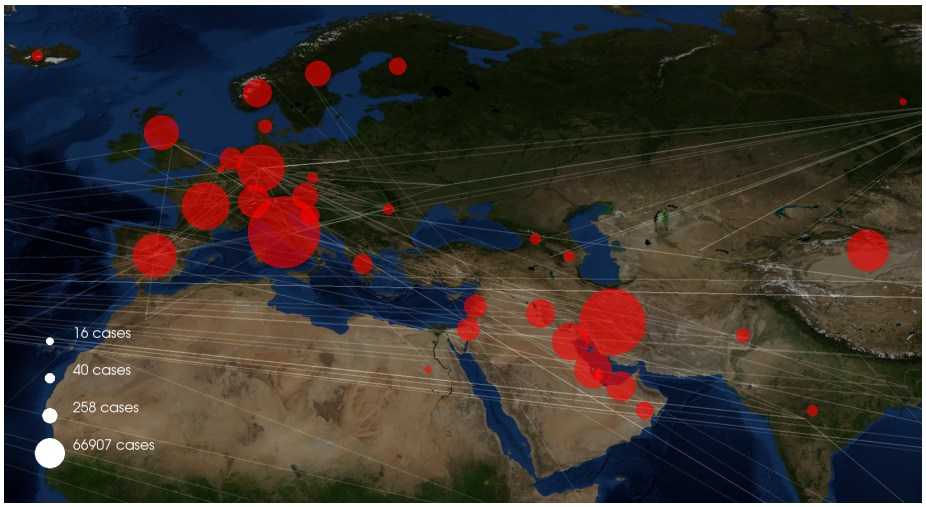

Figure 2: March 1st: Spread of COVID-19 in countries with high migration

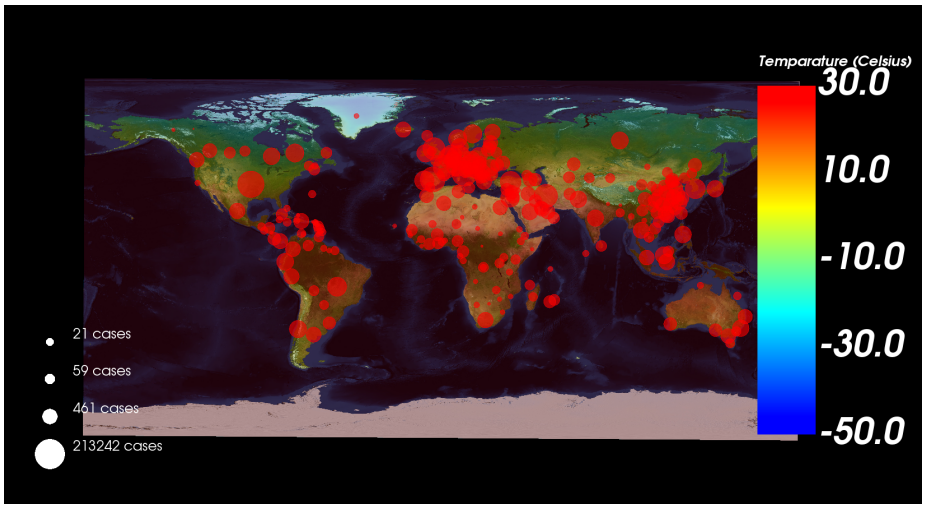

Figure 3: April 1st: Spread of COVID-19 and minimum temperature heatmap

The next set of screenshots show more clearly the spread of COVID-19 over time, along with recoveries and deaths. The chosen heatmap is maximum temperature, with migration data also shown so that connections with the factors can be seen:

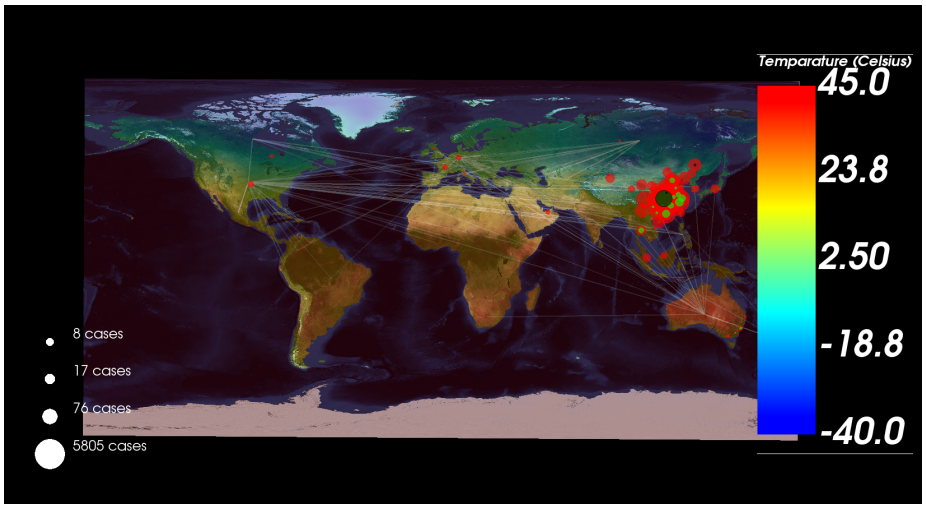

Figure 4: February 1st: COVID-19 data, with temperature and migration

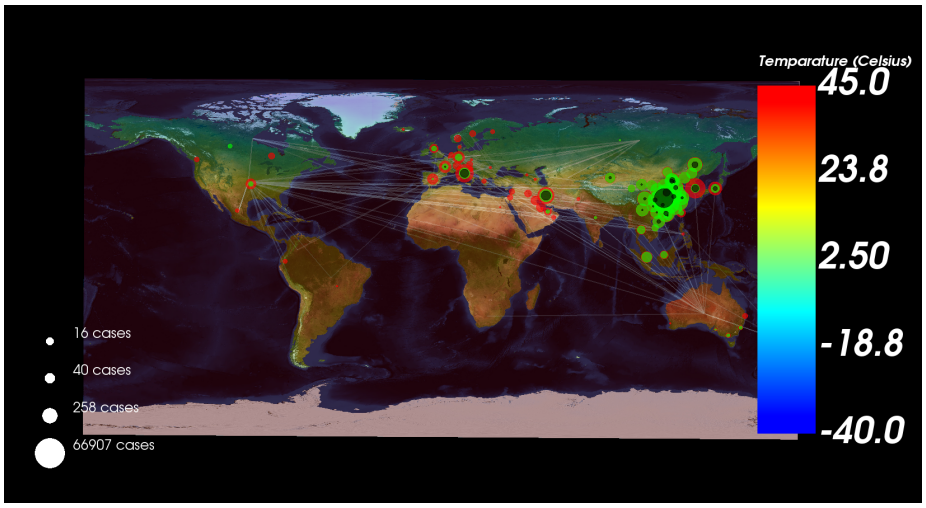

Figure 5: March 1st: COVID-19 data, with temperature and migration

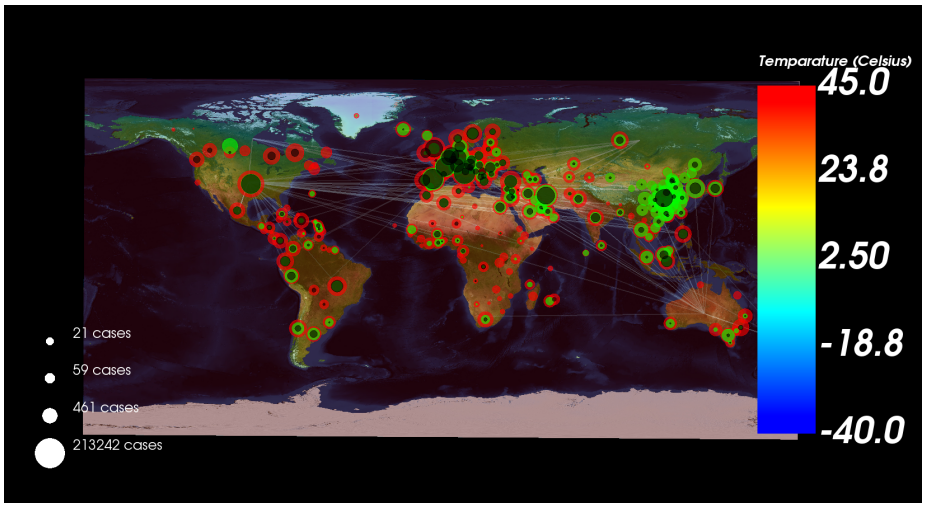

Figure 6: April 1st: COVID-19 data, with temperature and migration

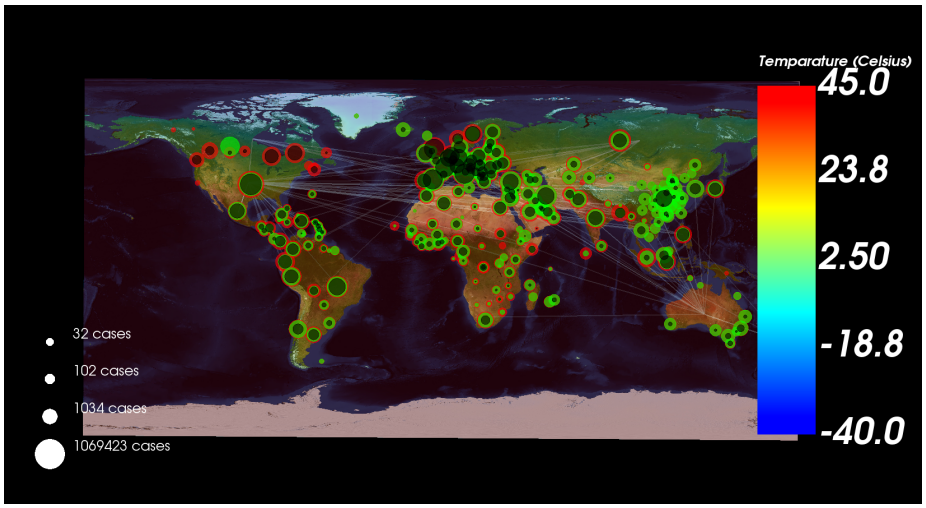

Figure 7: April 30th: COVID-19 data, with temperature and migration

6 Conclusions

From Figure 1, there seems to be some connection between population density and the spread of COVID-19, as can be seen in how Italy and Iran, two densely populated countries, are impacted more rapidly early on compared to less densely populated ocuntries like Sweden. However, the connection is not too strong, as some densely populated countries such as Poland are not impacted as much compared to Italy.

From Figure 2, there is definitely a stronger connection between migration and COVID-19 compared to population density. Countries with lots of immigration in Europe, such as Italy and France, are impacted much more heavily compared to those with little, like Sweden, or insignificant migration, like Poland.

From Figure 3, there appears to be a larger concentration of COVID-19 cases in countries with minimum temperature around 5-15 degrees Celsius. While the temperature does correlate with countries that actually have the capacity to test for COVID-19 on significant scale, there are some developed countries outside of this range that have notably less cases, such as Australia, which is warmer, and Sweden, which is cooler than the 5-15 degrees Celsius range. This data would seem to suggest that COVID-19 may spread better in the mild temperatures that are seen in Spring and Fall, as opposed to the colder and warmer temperatures seen in Winter and Summer.

From Figures 4-7, we can clearly visualize the spread of the virus over time. In Figure 4, we can see that China was hit very hard by the virus early on, as it has very many infections and very few recoveries. It is also clear that the virus had just started spreading to other countries outside of China around February 1st. In figure 5, after a month of spreading, we can see that the virus has spread through much of Europe and the Middle East, and that China has started to get the spread under control as there are many more recoveries. In Figure 6, after another month of spreading, the virus has engulfed the entirety of Europe and spread significantly on every populated continent. The United States appears to be the new leader in the amount of cases. In Figure 7, after about another month, we can see that the virus hasn't spread to many new countries, but there are many more overall cases. However, it appears that many countries are starting to recover, especially countries in Europe.