F. Quellec, L. Strauss, A. Ladoy

Process Book

Problematic

As the pandemic spreads around the world and population containment accelerates, social networks and search engines provide a window for people to learn and share about the virus. Mediatization of the pandemic shapes population's reaction to the virus, providing a support to rapidly share good practices about virus prevention but also a support for fake news spreading which could increase population anxiety.

Despite the fact that there are a multitude of interactive maps available on COVID-19, the vast majority only depict the spread of the disease. In our research, we wanted to relate COVID-19's spatial distribution with population reaction, defined here by the amount of information on both social networks (i.e. Twitter) and search engines (i.e. Google Trend).

Our main hypothesis was that the population response across European countries could be spatially heterogeneous, and not necessarily follow the spread of the pandemic. Indeed, although all European countries are now affected by the novel Coronavirus, it can be seen that even neighboring countries have employed different policies to fight the pandemic. Therefore, we could expect different population response too.

First step: Data collection

At least three datasets were necessary to answer to the problematic described above: data about Covid-19 prevalence, geolocalized activity on both Twitter and Google related to Covid-19.
We also had to collect auxiliary data, such as the administrative boundaries of European countries and their regions, as well as statistics (by country) that could be correlated with COVID-19.

Challenges
In order to be able to visualize these three pieces of information on an interactive map, we had several constraints: to have geographical data, with the same update frequency and covering the same time period (at least one month). The decision to use Twitter and Google Trends data was strongly dictated by their accessibility, their standardization for the whole study area, and the possibility of having geographical data, with daily temporal resolution. The fact that Google and Twitter are not available for all countries in the world (such as China) led us to restrict our study to European countries. We also had to narrow our analysis for February 19 to March 10, which was the overlapping time period between the three datasets we mined. As a future work, we will expand the time period.

A complete description of the data used in this project is presented below.

Name	*Description*	*Source*	*Spatial resolution*	*Temporal resolution*
COVID-19	Time series daily summary of reported Covid-19 cases and deaths (csv)	CSSE COVID-19 Dataset, John Hopkins University	Country-level	Daily
Tweets	Tweets with Covid-19 mentions that are geolocalized (~¼)	Twitter API, Twitter dataset	User address	Daily
Google Trends	Google Trends score [0-100] representing the popularity of the search (with term “Coronavirus”) by country on a specific day	pytrends, Google Trends	Country-level / Regional-level	Daily
Europe countries	Administrative boundaries of European countries (geojson)	NUTS 2016, Eurostat	1:10 Million	
Europe regions	Administrative boundaries of European regions (geojson)	NUTS 2016, Eurostat	1:10 Million	
Demographics	Population by age group and sex (csv)	Eurostat, 2019	Country-level	Yearly
Gross Domestic Product (GDP)	Gross Domestic Product in current $US (csv)	World Bank, 2018	Country-level	Yearly
Health Expenditures	Health expenditures in % of GDP (csv)	World Bank, 2017	Country-level	Yearly
Internet users	Internet users in % of total population (csv)	World Bank, 2017	Country-level	Yearly
National responses to COVID-19	National response to COVID-19 classifies in three categories (i.e. State emergency, Partial confinement, Total confinement)	Ouest France, 18/03/2020	Country-level	Yearly

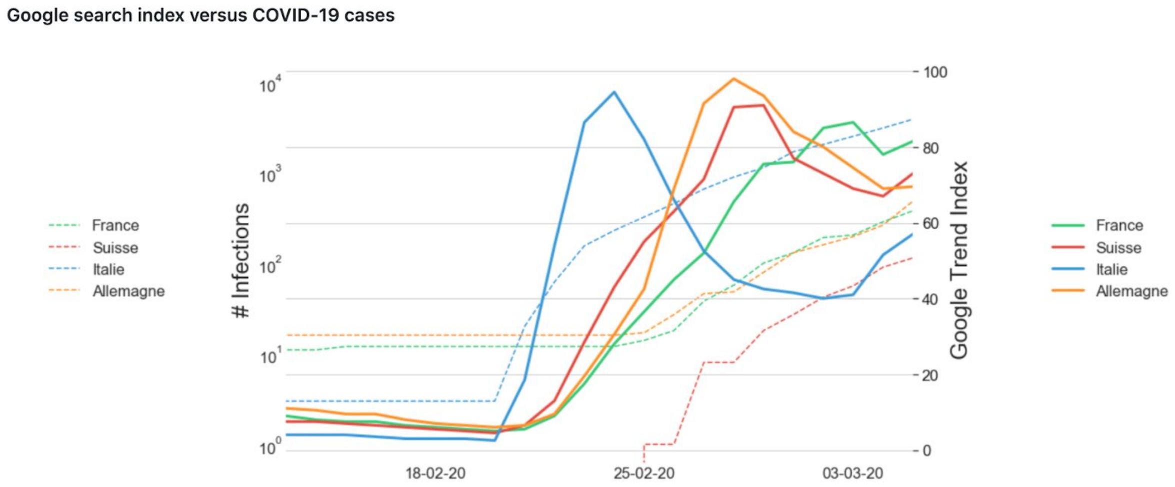
Second step: Data analysis

Cleaning and geocoding

We had to standardize the data to be able to use them easily on our website. Because our datasets came from various sources and file formats (.csv, .json, etc…), there were some inconsistencies in the name of the countries for example.
Geocoding was an important step to be able to match the Google Trends data with our geojson in order to build the choropleth map.
This was particularly challenging for the Google trends data at a regional level and we had to use the OpenStreetMap API to match the Google region name with the geojson regions.

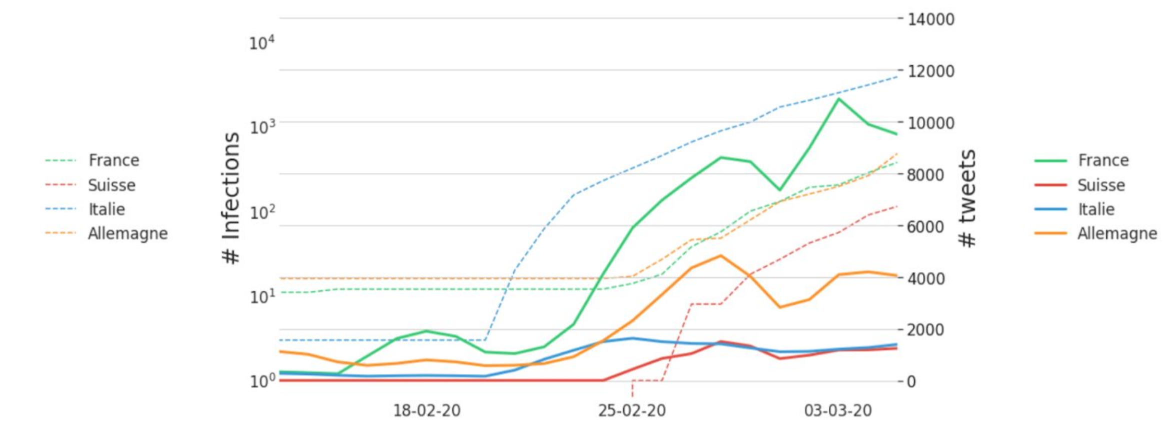
Descriptive statistics

As our objective is to study the association between the spread of the pandemic and the flow of information on social network and search engines, another step of data analysis was to run several descriptive statistics to better understand our data and see if a spatial visualization would be interesting.



For Italy, the google search spike coincides perfectly with the outbreak of COVID-19 cases in the country. It is also interesting to note that the countries surrounding Italy are seeing an increase in the number of searches a few days after Italy, which seem to be related to the spread of the virus in Italy and not only in their respective countries.

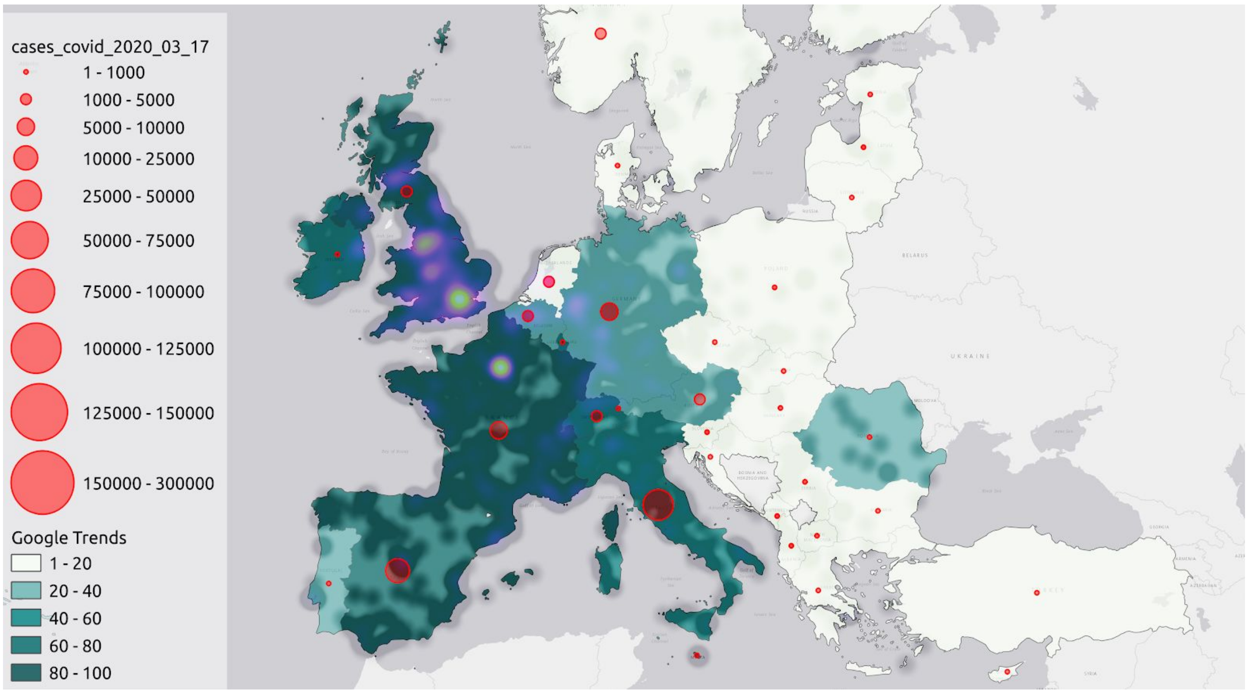
Number of tweets versus COVID-19 cases



With regard to the number of tweets about COVID-19, although trends and uses are different in each country, the fluctuations still seem to be directly related to the spread of the virus in each country. The two graphs presented above indicates a correlation between the reaction of the population on Twitter and Google, and the geographical evolution of the pandemic. These findings are encouraging us in the desire to explore these variables on an interactive map.

Third step: Data visualization

The last phase of our project was to build a web application using data visualization techniques that will summarize the information contained in our data in a efficient and esthetic way. First, we sketched the three variables with QGIS to make sure it was readable (see figure below). Some small changes have been made during the implementation and we describe each feature in the next parts.



Interactive map

This is the centerpiece of our visualization. Because we had a lot of data to display, we had to choose judicious visualizations so that we could overlay our different data layers and keep a readable result.

We chose three kinds of spatial visualizations described in Lecture 8 (Maps) that were good fit for our data and that would not overlay too much: Bubbles for Covid cases, heat map for Twitter data and choropleth map for Google Trends data.

Challenges:

- Find a color palette that looked good while highlighting the correlations but without having one particular type of data standing out.
- Make the application responsive with different size of screens (not fully done)
- Make the application compatible with several browsers (for now prefer Google Chrome)

Bubble map of COVID-19 reported cases

We chose to represent the number of confirmed cases (only where we have available data) with bubbles at the country's centroid. These bubbles scales logarithmically to avoid overloading the map for higher numbers.
They are displayed red with a white border stroke to help them stand out a little bit with respect to the heat map.
We also decided to make the number of cases appears directly in the bubbles for more clarity, the font size grows with the bubble scale.

Choropleth map of Google searches related to Covid-19 by country and regions

The countries are simply colored with the scale described above. When zooming on a country, we display the individual regions and color them using the corresponding Google Trends index, relatively to the country itself. This means that if Britain had the biggest number of search on a particular day compared to other French regions, its index will be 100 out of 100, if another region have half of this search volume, it's index will be 50. The same thing happens when we scale to european countries.
The colors scales linearly from a soft white to a blue from the official Google color palette.
The zoom feature was particularly difficult to implement since we are displaying 4 different layers of informations, the covid bubbles, the map, the heatmap and the legends and panel, all reacting differently to the zoom transformation.

Geographical heatmap of geolocalized tweets related to Covid-19

A geographical heatmap is a powerful way to summarize the information of thousands of points because it interpolates the density of points around a certain radius with a Kernel Density Estimation.

For the implementation, we were inspired by the simpleheat.js library.

One difficulty of this type of visualization is to define an optimal radius which will not cover entirely the map and will not give too much weights to individual points (only one tweet). We decide to use a radius of 10x10 pixels. A radius in geographical units would have been better but was difficult to implement especially with the rescaling when we zoom on a country.

As the heatmap is a canvas layer and the other data elements are svg, we had to draw the heatmap below the geojson to allow interactivity with the countries (e.g. zoom) and thus, a challenge was to choose an optimal opacity to see all the layers in the map.

For the colors, chose the Player yellow to Melmo palette because we wanted to keep the concept of “heat” map by using warm colors as the numbers increase. However, the heat map begins with a soft blue color for very low values to better blend with the rest of the visualization.

Space-time visualization

In order to travel through the dimension of time, we have a slider at the bottom of the page that changes the current date to display. We just had to be careful to update what’s needed when the date changes and handle the good time formats.

Right panel

On this panel we display all the data we have on the selected country. The graph displays both the number of cases and the number of deaths on a log scale. If we use a linear scale graph the number of deaths becomes to small comparing to the number of confirmed cases and the graph becomes useless. There was the possibility to have one graph for each data but we felt like it would be harder to see the correlations and the panel would become to large.

Left panel

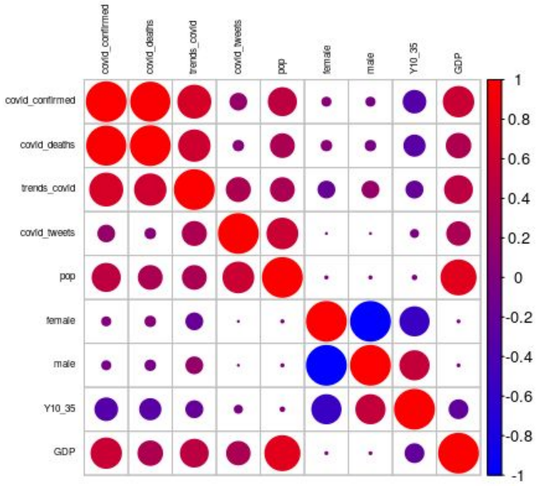
Top 10 countries

The top 10 countries with the most registered cases at the current date are displayed here. This allows the user to have a global view of the situation even when focusing on a country.

Solar correlation

The interactive map allows the user to explore the spatial distribution of the prevalence of Covid-19 cases, tweets and google searches.

At Milestone 2, we had already seen a strong correlation between the number of COVID-19 cases in a country and some of the country characteristics such as GDP and total population. The correlation between our three main variables and several covariates was represented using a heatmap in R (see figure below).



In our final project, we made several changes: first we added several covariates (list is provided below) and we decided to use a solar correlation plot instead of a heatmap that we implemented using D3.
We chose to use a Solar Correlation Plot because we are mainly interested in the relationship between COVID-19 cases, and this type of visualization represents an intuitive way to explore the influence of several covariates on an output variable and also the intercorrelation between the explanatory variables (Zapf and Kraushaar, 2017). For the implementation in d3, we were inspired by the Python script implemented by Zapf.

We can interpret the plot this way: the output variable (in our case, the total number of Covid-19 cases by country) is the sun, at the center of the solar system. Orbits represented by radial line around the sun corresponds to the level of correlation (Pearson correlation coefficient going from 0 to 1) with an increasing correlation towards the sun. Each explanatory variable is a planet orbiting around the sun with the distance depending on the level of correlation with the output variable. The planet appears in green if the correlation is positive and red if the correlation is negative. If two explanatory variables are highly correlated (we put a cut-off level of 0.7), the variable with the strongest correlation to the output variable becomes the planet, and the others its moons. Such variables are surrounded with a grey stroke. Writing the names of all the variables made the plot unreadable, so we implemented a “mouse over” to display both the name of the variables and their role (sun, moon, planet with a moon) in order to facilitate interpretation.

It should be noted that all the main variables (COVID-19 cases, COVID-19 deaths, Google Trends and Tweets) were aggregated beforehand by country and over the time period February, 19 - March, 10 to be able to compute the correlation with different country characteristics.

List of variables in Solar Correlation plot:

- Total number of reported COVID-19 cases for the time period (output variable)
- Total number of confirmed COVID-19 deaths for the time period
- Gross Domestic Product in current $US

- Internet users in % of total population
- National response to COVID-19 classifies in three categories (i.e. 0:State emergency, 1:Partial confinement, 2:Total confinement)
- Health expenditures in % of GDP
- Total Google trends score for the time period
- Total number of tweets for the time period
- Number of neighboring countries
- % of elderly population
- % of young population

Peer Assessment

What	Who
Data collection	François / Lucas / Anaïs
Data analysis	François / Anaïs
Bubble Map	Lucas
Choropleth	François
Heat map	Anaïs
Time slider	François
Right Panel	Lucas / François
Left Panel: Top 10 countries	Lucas
Left Panel: Solar Correlation plot	Anaïs