Process Book: Covid-19 in South-Korea

Lucas Eckes, Anton Soldatenkov & Frédéric Bischoff

Coronavirus is the hot topic of the moment and mobilizes the skills of data-scientists like never before, and thus not only for describing the spread of the virus, but also for modelling future and even for helping researchers to develop drugs.

On our scale, by exploiting a Korean dataset (https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientInfo.csv), published by the KDCC (Korea Center for Disease Control & Prevention), we want to provide some visualizations allowing to observe the impact of parameters that have been less analysed so far. Indeed, we would like to provide some representations that shows for example, if some characteristics of a city, like population density, the proportion of persons in the elderly or even the number of schools may play a role in the virus propagation. In the same fashion, the hypothesis that the weather has an impact and in particularly the temperatures is often mentioned. Again, we would like to provide some visualisations, to see if we can confirm or not some of these assumptions.

In this document, we present in details the provided visualizations and path that carried us to realize them that way. For better clarity, we will present our work in three main parts: **A first part** is dedicated to visualisations giving a precise overview on the virus spread in South-Korea since January. **The second part** focuses on the impact of city morphology on virus propagation, and the **third part** presents the visualization aiming to show any link between weather and the virus. In a last part, called **peer-review**, each of us summarizes the work done by himself and the other members of the team.

I. Spreading of Covid-19 in the South Korean population

First interactive map:

Motivation: The goal of this part was to explore the dataset for different provinces and social groups in South Korea. First, it was important to convey **the extent and severity of COVID-19**. For that, two aspects were important, the **time** to show the speed of transmission and the **fatality rate** and **cases** in the population to demonstrate the severity.

Design and achievements: To render this, a dynamic map was created with respect to time and with colors related to the concentration of cases among population. Two maps were provided, a map consisting of **provinces** and a lower level administrative division consisting of **municipalities**. It is possible to **zoom** and **hover** on regions to have the detail of cases. The color palette was created using chroma.js between a specific green and red that maximizes the lightness and hue difference. **Special events** are displayed on top of the time slider as desired. (1.A) The spread in the country is stunning especially after 18 February when the "patient 31" participated at the Shincheonji Church gathering. The population was then separated according to their age group and gender. It is possible to **change the group** and the **map will update accordingly**. (1.B) Initially, a curve of total cases was proposed. But after consideration, a **bar plot** will display better the differences among groups and regions than multiple curves. (1.C)

Challenges: The struggle was to compute the fatality rate among groups. Before 2 March, it was not possible to get consistent death data. Moreover, the Daegu Metropolitan City Hall does not disclose all the information about more than 6000 individual cases resulting in case mismatches. That's the reason why on 2 March there is a huge climb in the number of cases, therefore we decided not to display the bar plot before this date. However, that doesn't minimize the huge increase of new cases observed after 18 February.

Another initial choice was to select regions according to their number of schools. However, since we didn't have access to death data for municipalities, the map will not change apart of some regions with null values. Furthermore, selecting regions depending on schools or elderly ratio will probably overlap with the second part.

The transmission rate was not calculated as it requires more complex data than just the number of cases and the number of deaths.

Second interactive map:

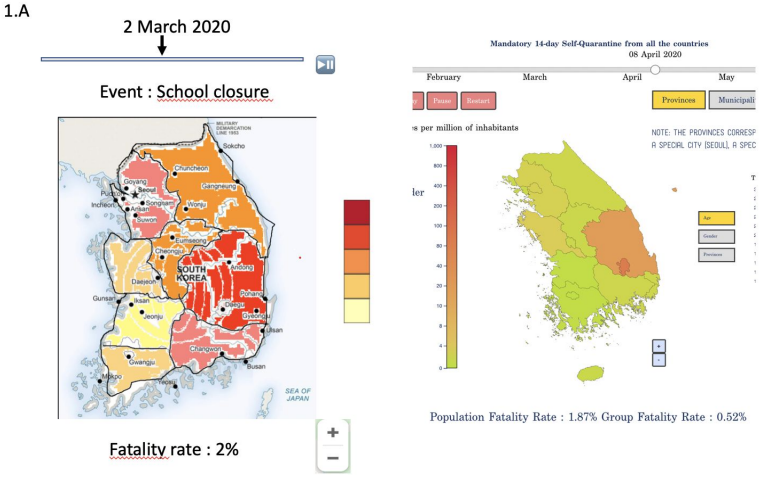
Motivation: The first interactive map presents the disease as part of population concentration but doesn't unveil the daily changes or the cases at the **individual level**. This map will focus on **daily new cases** and concentrate on their **causes**.

Design and achievements: For the daily new cases, the scaling was made using **circles** filled with the same color. The area of the circle will depend on the number of daily new cases (2.A). The causes were **listed** from the top cause to the cause with the less number of cases. An **emoji** was downloaded for each cause to make it more easily identifiable. The total cases were written next to the cause (2.B). Again, it is possible to filter the age and gender while the map and causes will automatically be updated.

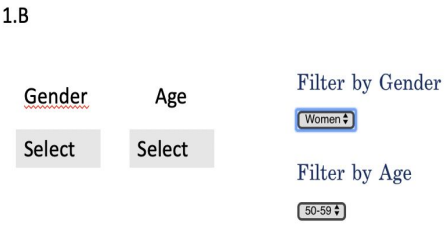
Challenges: It was better to take a scale linear with respect to the area and not the radius since the data has a large domain from 0 to 2000 cases (2 March). However, the problem isn't solved as dozen of cases are represented by too little circles and 2000 cases cover all the Daegu province. A better scale to implement may be d3.scaleSymlog.

The selection of regions and the curves representing the cases and deaths were not implemented for the same reasons as mentioned above.

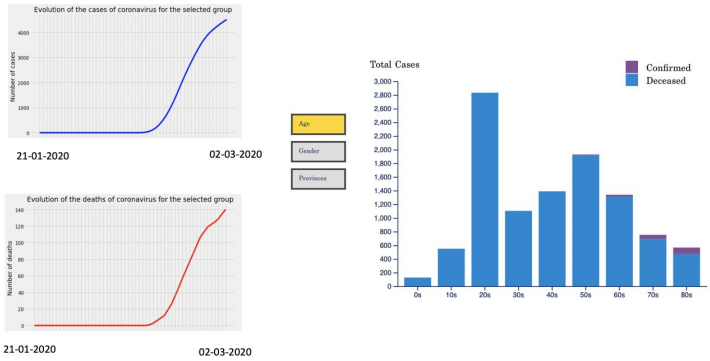
1.A



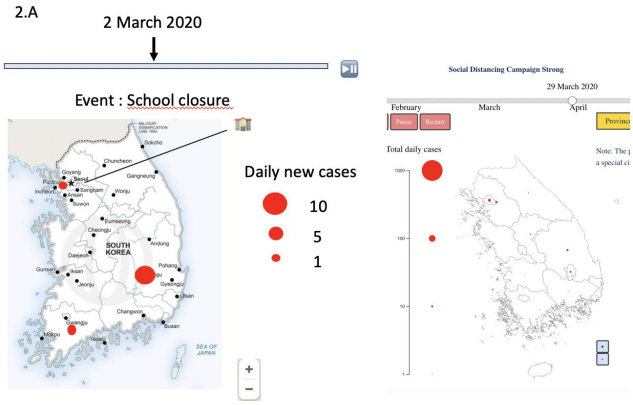
1.B



1.C



2.A



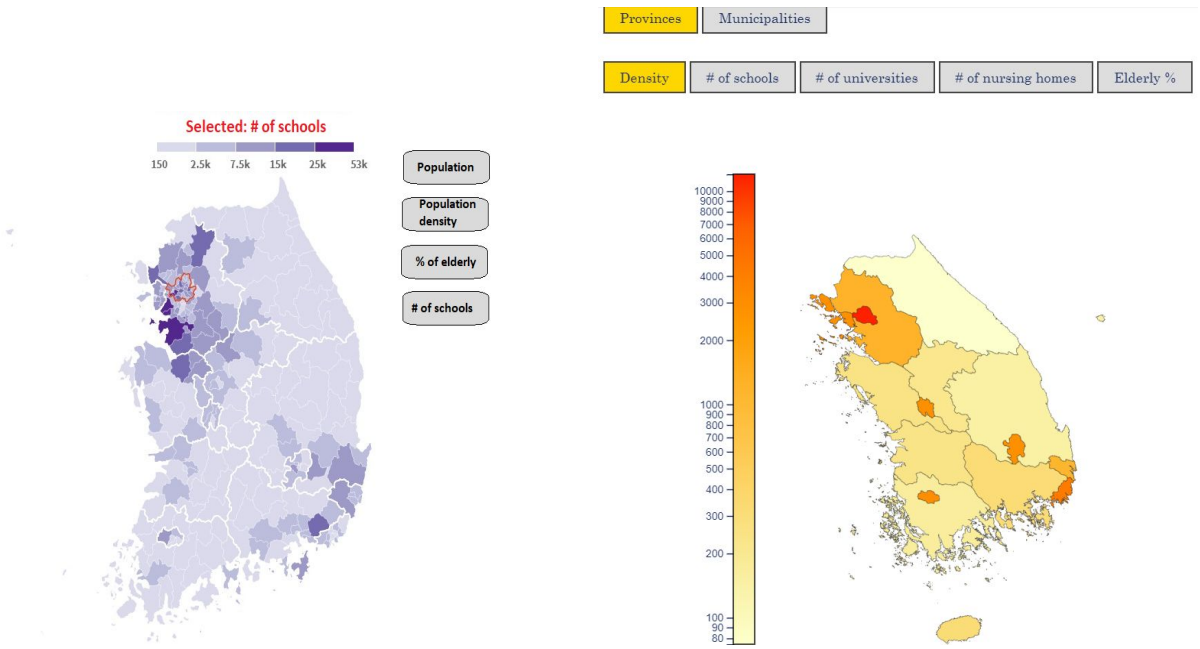
2.B

Milestone 2 vs final version: For each of the figures above, the final version of the website was compared to Milestone 2's sketch. The left hand side of every figure corresponds to the initial sketch whereas the right hand side corresponds to the final version that was implemented. It is possible to observe the changes and progress of ideas while the guideline remains the same.

II. Role of city morphology

In this section we provide information about the provinces of Korea that can be assumed **constant within the considered time frame**. As we realized after starting working on this part, not all data was present in the original dataset (e. g. population density), so we looked for reliable and up-to-date sources (National Statistical Office of the Republic of Korea, Ministry of the Interior). The resulting **map** depicts different aspects of city morphology (population density, number of schools, universities and nursing homes and the percentage of elderly population) which can be changed using the **buttons** on the top. The selected feature is highlighted in yellow. It is worth noticing that the **color scale** on the left side of the map varies adapts on the selected feature: it is **logarithmic** for data with a big difference in orders of magnitude and **linear** otherwise, its lower and upper bounds are derived from the data. A **tooltip** with the name of a province and the exact value of the selected feature is shown when the mouse covers one of the provinces on the map. We also tried to find out if any of the considered aspects has an impact on the virus spread. To do this, we calculated **Spearman correlation** between the features and the number of cases. For features that depend on the size of the population (e. g. population density, % of elderly people) we considered cumulative prevalence (number of cases per fixed number of inhabitants) instead. The results are shown in the **table** below the map. We can observe a high correlation (›0.7) between the number of nursing homes and the number of cases. Indeed, we can find news about senior-care facilities being the centers of the outbreak in South Korea (https://www.scmp.com/week-asia/health-environment/article/3075937/coronavirus-nursing-homes-emerge-south-koreas-new).

The correlation with the number of schools and universities is also significant but we could not find any reliable evidence that they influence the spread of the virus more than any other public places. Regarding our visual implementation, we decided to change our color scale to be in accordance with other maps on the page



Milestone 2 vs final version: The concept remains the same, we changed the location of the color scale and its style and moved the buttons on the top.

III. Impact of the weather

In this part of the website, we provide some tools to **explore any correlation** between weather data and the propagation of the virus. For that, we simply calculate the **Pearson correlation** between the chosen **meteorological parameter** (Temperature, Humidity, Rain, Wind) and the **number of daily new cases**. But of course, the main difficulty is that the **weather at a given day**, may have **an impact** on the number of new detected sick persons **only few days latter**. For that, we let the ability to create a **time lag** between the two signal : a time lag of two days means that we calculate the correlation of the number of new sick persons with the weather data from two days before.

The realized plot **(Fig.1-realized)** allows to manually choose a delay (from 0 to 40 days) with a **slider** and observe the variations of the delayed signal to align « virtually » the weather who may had an impact on the number of new sick person few days latter. Instead of just displaying the correlation score between the curves as we wanted to do at the beginning **(Fig.1-expected)**, we dedicated a vertical axis with a **moving point** two see more visually how strong is the correlation. This plot let the ability to choose on which parameter we want to explore any correlation (Temperature, Humidity, Rain, Wind) with the help of **four buttons**. By staying on a given point you get its **exact value**, and if needed you can **zoom on a given** period by selecting desired period of time, while keeping the time lag. For this plot, it's for example **interesting to see** that for a delay between 14 and 16 we get struggling **correlation scores around -0,8** between temperature and nb of patient, this would mean that increasing temperature would have an negative impacts on the virus, which is in accordance with what experts are saying: "high temperatures limit the virus propagation".

The next figures of this part is a heatmap that summarizes the correlations explored with previous graph to get an overview of the correlations. The figure obtained is in accordance with the expected one, as you can see on **Fig.2**.

Lastly, we wanted initially to realize a third plot for this part, it should have shown the time-lagged correlations per province and would allowed to see if an effect observed in one province is the same in another one. But it appeared that for some province data were missing for some days, and so, doing a time a lag of one day would have not made sense if there are some holes in the time series, as a consequence, we worked on averaged values and the daily sum of new patient per day for all South Korea.

Fig.1: Weather graph n°1, differences between expected and realized figure

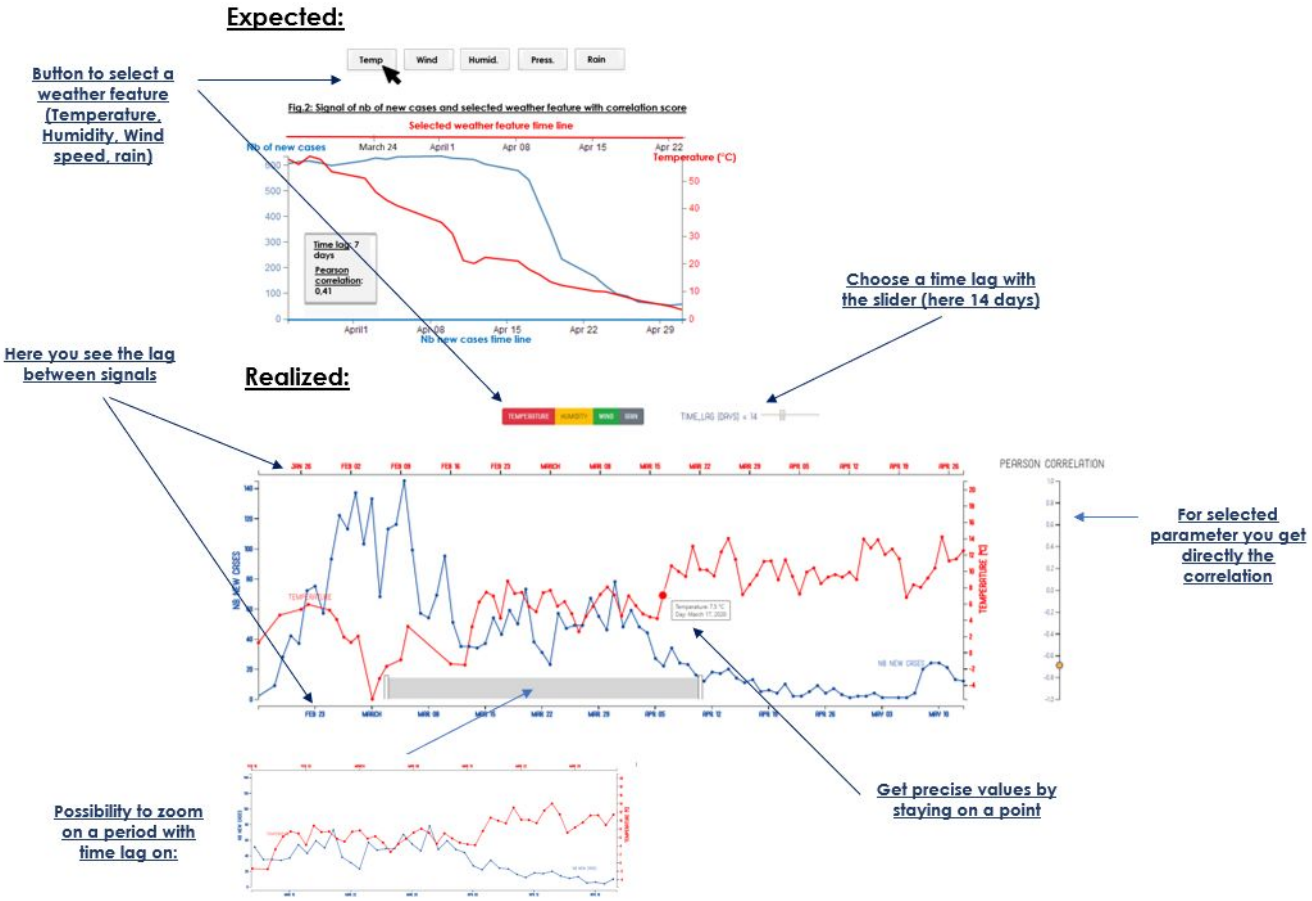
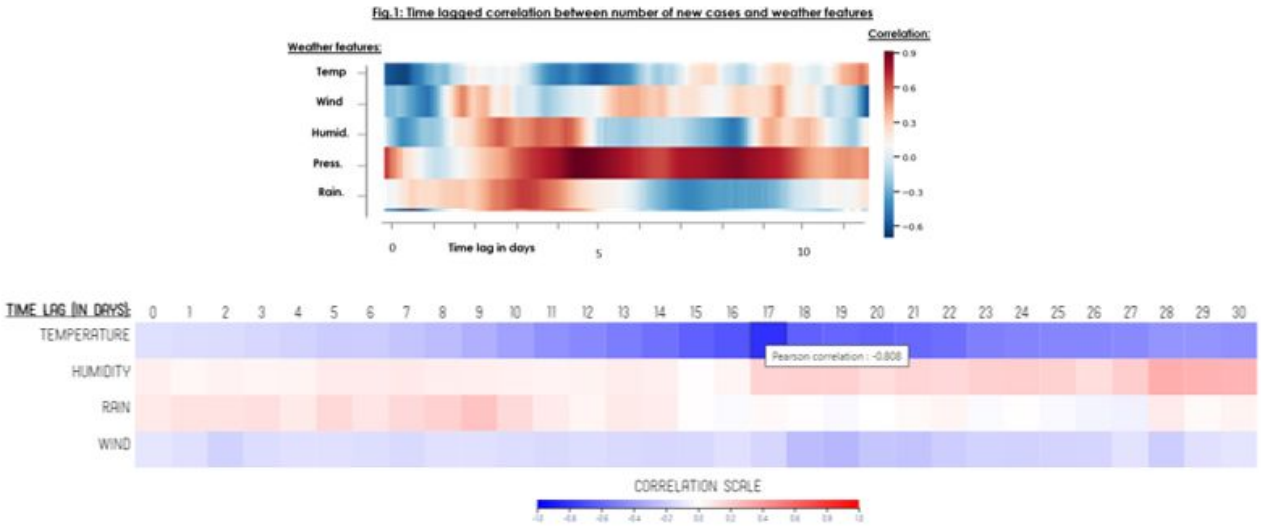


Fig.2: Weather graph n°2, differences between expected and realized figure



For humidity and rain the correlations are always under 0,3 and not significant enough to draw any conclusion. At the opposite it's interesting to see that for a delay between 14 and 16 we get struggling correlation scores around -0,8 between temperature and nb of patient, this would mean that increasing temperature would have an negative impacts on the virus. This is not surprising because all specialist says that hot weather decreases the spread, but it's interesting to see that it lasts approximately 2 weeks before observing the impact of temperatures. Concerning wind, we observe an interesting correlation around -0,4, quite smaller than the correlation for weather and less reliable. An hypothesis could be that the virus that can survive 3 hours in the air could propagate faster in the air with wind, but that does not seems to be the case. Humidity should be favourable to virus but we see here that ...

IV. Peer review

In this part, each of us describes on which parts of the projects he focused and what the others did.

Frédéric Bischoff:

I realized the entire work on the "weather" part, that means the two visualizations for this part:

- **1.** the interactive two axis scatter plot that gives the possibility to time lag the evolution of one weather parameter from the number of new sick persons, in order to explore any correlation and
- **2.** the heatmap summarizing the correlations

Naturally, I also did the work around this plots, that means, the description and interpretation of these graphs.

I also realized the background of the website and choose the style, the fonts and added a sidebar to switch easily from one part to the other. I worked on the redaction of the introductory text and the problematic for the website and this process book. Finally I also had to realise entirely the video describing our work (this task that was dedicated to Anton originally)

Lucas realized the graphs of the first descriptive part and Anton realized the map of the second part about the impact of city morphology and posted the video on Youtube.

Lucas Eckes: I realized the first part of the website corresponding to the spread of coronavirus in South Korea. I implemented all the tools present in the first and second interactive maps. In addition to that I made the data cleaning and search for the TopoJSON files. For my part, I needed to import wikitables for municipalities population and had to download a csv file from internet giving the population by age and gender in South Korea. Anton created the second part and did the Youtube video . Frederic made the last part, took care of the website background style and the introductions.

Anton Soldatenkov: I created the "city morphology" part of the website. I mainly worked with static (not changing over time) data. It turned out that a lot of important data was missing or did not make any sense, so I found some reliable data sources to complete it. I created an interactive map of Korea provinces and municipalities that shows different city morphology features (in particular, population density, number of schools, universities, elderly homes and percentage of elderly people). I also calculated Spearman correlations of these features with the number of cases, added a small table to the section and looked through news websites, newspapers and scientific articles to confirm or reject my hypotheses.