COVID-19 in Poland - Project Overview

Bartosz Żbik

December 3, 2020

1 Brief Overview

The aim of the project is to analyse the outbreak of COVID-19 in Poland. Source data about the disease aren't easy to get in Poland. Data I managed to get are extracted from various resources by Michał Rogalski. Data can be downloaded here. I import the data, change the data format to *.csv* and than analyse it using pandas library for python. The output is than plotted using matplotlib and plotly libraries for python, and finally uploaded on the school server.

2 Description

2.1 Data Source

Data I analyse are available here thanks to Michał Rogalski. I download those spreadsheets in *.csv* format, extract some data and change the data structure to make it easier to analyse. The main reason I format the data in a different way is that I have previously used different data sources.

In the past I have used data from this github repository, thanks to Vaclav Rut from Apify team. The source for this repository is the government webpage. Data can by dowloaded from this webpage and the latest record from this webpage.

2.2 Extracted Data

To extract and analyse the data I use the pandas library for python. Some early data ware updated more than once a day, but it is no longer supported, so I decided to resample the data, so that I have one record per day. Beside the total number of cases and deaths i compute the daily number of new cases and deaths. I also compute the rolling averages of new cases and deaths taking different window sizes. For now those are 3, 7 and 14 - day averages. I also compute some custom parameters like local trend and try to predict the future outbreak using a simple model. Usually all data I extract are computed for the

whole country as well as for each region on its own. The analysed data are saved locally to .csv format, wich I prefer over .json.

Stored columns are for now:

Table 1: Data I extract and store in data files

Column Name	Description	Important Notes
region	"Województwo", no polish letters	Only in region data
date	One record per day	1 day resolution
infected	Cumulative infections	Imported from web
dead	Cumulative deaths	Imported from web
new	Daily number of cases	Computed
new_deaths	Daily number of deaths	Computed
new_3	Rolling 3-day average	Computed
new_7	Rolling 7-day average	Computed
new_14	Rolling 14-day average	Computed
new_deaths_7	Rolling 7-day average	Computed
new_deaths_14	Rolling 14-day average	Computed

2.3 Why Those Data?

To analyse data, I use rolling averages with window sizes 7 and 14 days. The data has clear weekly oscillations, because of that I decided to use windows with size $k \times 7$days, $k \in \mathbb{Z}$. Using something like 10-day average is not working correctly because of the weekly oscillations, wich I analysed manually. 21-day windows don't smooth the data much better than 14-day windows, but the 21-day average changes very slowly compared to 14-day average (wich is already a week late against the situation), thats why I don't use it.

To analyse weakly oscillations I use a 3-day average and compare it to the 7-day average as some kind of reference. Weekly oscillations are quite well visible on daily cases, but the daily number of deaths isn't big enough to see the oscillations clearly.

2.4 My Output Data Format

The data I extract are usually shown through plots. I use the matplotlib library to plot most of the data. I try to keep a consistent plot format for all plots so, that the viewer doesn't need to read the legend every time, but it my change in the future to make some plots easier to read.

Some parameters are nice to show on a map. To make those maps I use the plotly library. I use an outline map of Poland wich I downloaded from this link, thanks to Filip Stachura, who uploaded it on his github.

For now plots and maps are saved as interactive plots in .html format or are exported to .png pictures. The second option is usually used if the .html would be to big.

3 Computed Parameters

3.1 Local Trend

Local trend is intended to show the current situation in each region. Currently it is computed using this equation

$$\text{local trend} = \frac{\text{"new\_7" for today}}{\text{"new\_7" for week ago}} - 1.0 \ . \tag{1}$$

If local_trend > 0 than the situation is becoming worse relative to what was observed before. I try to make this parameter stable, but it makes it also react slower to the situation.

4 Modeling

4.1 Prediction Curve

To compute the prediction curve I need to estimate the current reproduction rate of the virus. My model assumes, that

$$\text{new cases}(t + \Delta t) \propto \text{active cases}(t) \tag{2}$$

and that

$$\text{active cases}(t) = \sum_{t-n}^{t} \text{new cases} \ , \tag{3}$$

where n is the number of days for wich the case stays active. In my model, I assume $n = 10$. From equation (2), fot $\Delta t = 1$ day, I compute the "daily reproduction rate", wich is the number of new cases, generated by each active case every day

$$\text{daily reproduction rate} = \frac{\text{new cases (today))}}{\text{active cases (yesterday)}} \ . \tag{4}$$

This parameter is strictly related to the assumption that $n = 10$. To convert it to something widely used I need to multiply it by n

$$R_\varepsilon = n \times \text{daily reproduction rate} \tag{5}$$

where R_ε is called R effective and it is the number of new cases generated by each case. As far as I know it's not the R_ε we can find in bibliography, because it's responsive to the

situation in Poland, such as opening or closing schools. On the other hand it quite well describes the current situation.

For now the prediction curve refers to the rolling 7-day average. I count the average "daily reproduction rate" from the last 7 days, and than using the above model I try compute the predicted future averages. The error lines are computed using the same model but for "daily reproduction rate" $\pm\, 3 \times$ standard deviation.

4.2 Modeling Based On Hospitalizations

I have access to the number of cases that are currently hospitalized. I do not know the number of new cases that are hospitalized or the number of cases that finished their hospitalization period. In my model I assume that the average hospitalization period $T_H \approx 16$ days, which I assume based on [1]. The population sample described in the article is not similar to the population structure in Poland, but all parameters where measured very accurately compared to other sources. I then use a following approximation

$$\text{new hospitalizations (avg)} \approx \frac{\text{currently hospitalized}}{T_H} \ , \tag{6}$$

where *new hospitalizations* is approximately the average numer of new hospitalizations (cases which started their hospitalization period) in last T_H days. Assuming that *new hospitalizations* $\propto$ *new cases* it's possible to approximate the number of *new cases* based on *hospitalizations*. The advantage of doing it indirectly is that it is less dependent on the way the government is testing for *COVID-19*. Combining whats said above we get the following equation

$$\text{new cases (avg)} \approx \frac{\text{currently hospitalized}}{T_H \cdot H\%} \ , \tag{7}$$

where $H\%$ is the percentage of cases that required hospitalization in the number of cases. $H\%$ depends on the population structure and the way the population is tested for *COVID-19*. The second factor is more relevant in case of Poland. In my predictions I use $H\%$ fund in the article [1]. Because of that the number of cases found this way will not be similar to the measured new number of cases. It will not be the number of all cases that occurred either. The reason for that is the difference in the population structures between Poland and the aircraft carrier mentioned in [1].

I then use the same model to predict the future outbreak as I applied to the number of new cases. It is important that there is a time shift between new cases and new hospitalizations, but I have not yet found the shift value so I use $\mathit{shift} = 0$.

References

[1] Matthew R. Kasper et al. “An Outbreak of Covid-19 on an Aircraft Carrier”. In: *New England Journal of Medicine* 0.0 (0), null. DOI: 10.1056/NEJMoa2019375. eprint: https://doi.org/10.1056/NEJMoa2019375. URL: https://doi.org/10.1056/NEJMoa2019375.