Considerations on how data is displayed and interpreted in the COVID-19 pandemic

Beatriz Stransky[1,2], Guilherme F. Araujo[2], Marcus A. Nunes[3] & Sandro J. de Souza[2,4,5]

1- Department of Biomedical Engineering, UFRN
2- BioME, IMD, UFRN
3- Department of Statistics, UFRN
4- Brain Institute, UFRN
5- Institutes for System Genetics, West China Hospital, Sichuan University

Corresponding Author: Sandro José de Souza, sandro@neuro.ufrn.br

Abstract

In situations with high social impact, like the COVID-19 outbreak, the form on how data is displayed and interpreted has a critical role in several issues related to public perception and actions. Here we developed a platform that generates and display epidemiological data related to COVID-19 in different visualization formats. The combined use of different data visualization formats allows users to have a better understanding of the population dynamics of SARS-CoV-2. Data and graphs can be found at https://bioinfo.imd.ufrn.br/COVID-19.

Introduction

As the SARS-CoV-2 spreads rapidly around the globe, it is extremely important that epidemiological data can be retrieved in a fast and reliable way [1]. Such data are crucial for feeding mathematical models (like in [2-3]) that can predict the course of the pandemic and define actions from public and private institutions. At the same level of importance is the use of such data, together with visualization tools [4], to inform the public about the course of the pandemic and engage them in actions that will minimize the spread of the virus, like social distancing [5].

It is our understanding that this last component has been neglected by both official government institutions and the overall press. This seriously undermines the whole effort of better informing the population about the crisis and affects the rate of engagement of the public regarding actions to minimize the consequences of the pandemic.

Here, we discuss different types of visualization formats that are more effective in describing the population dynamics of SARS-CoV-2. A web page is available to make the data and graphs public. This portal will be continuously updated.

Methods

Epidemiological data in CSV format was obtained from the initiative “Our world in data” (http://ourworldindata.org) from the University of Oxford [6]. That includes: number of cases per day, number of tests per day and number of deaths per day. Graphs were designed using ggplot2 [7] and Matplotlib [8].

Results and Discussion

By examining the broad press coverage of the COVID-19 pandemic, one issue that called our attention was the poor representation of the epidemiological data. To communicate information clearly and efficiently, data visualization merges statistical analysis with design, using visual elements like charts, graphs, and maps to visually communicate a quantitative message. This is a major element in data science. Effective visualization provides an accessible way to identify trends, outliers, and patterns in data, helping the users to reason about data and evidence [9].

Take, for example, the most common way of representing the number of COVID-19 cases identified in a given country (Figure 1). The reported values are simply the cumulative number of cases along a time window (usually in days). The way the data is displayed does not allow one to discriminate between several parameters that are affecting the shape of the curve. With the example shown in Figure 1, one cannot precisely define if the growth is exponential because, among other things, the number of tests is not taken into account. The growth suggested by the curve may be due to a larger number of tests being performed more recently. What is surprising and concerning is the fact that the official reports from the Minister of Health of most countries also lacks a more rigid statistical evaluation of the data.

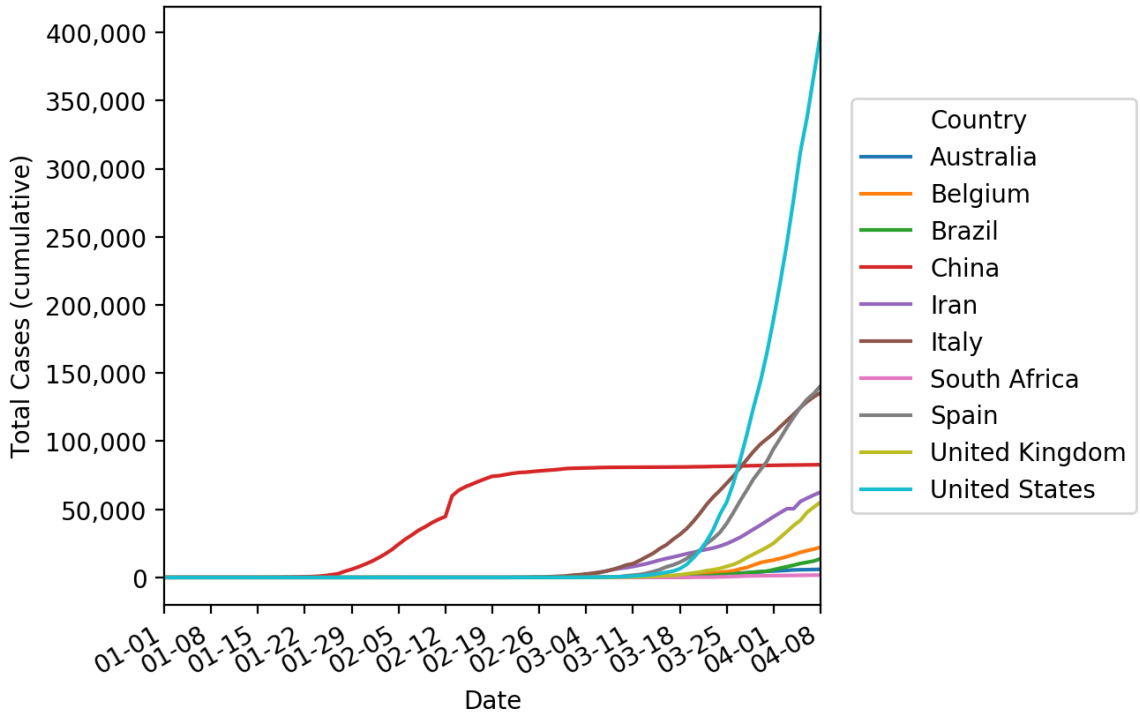

Figure 01: Cumulative cases for ten countries, from January 1st to April 8th.

A more correct way of depicting the population dynamics of the virus by plotting the number of new tested cases (in the Y axis) by the total (cumulative) number of cases in the X axis (both in the logarithmic scale). Figure 2 shows such a graph for the same ten countries, as in Figure 1, during the COVID-19 pandemic. While this type of graph is extremely useful in evaluating whether the virus in a given country is going through an exponential growth, there are several caveats, the most seriously being that time is not represented (only through an animation).

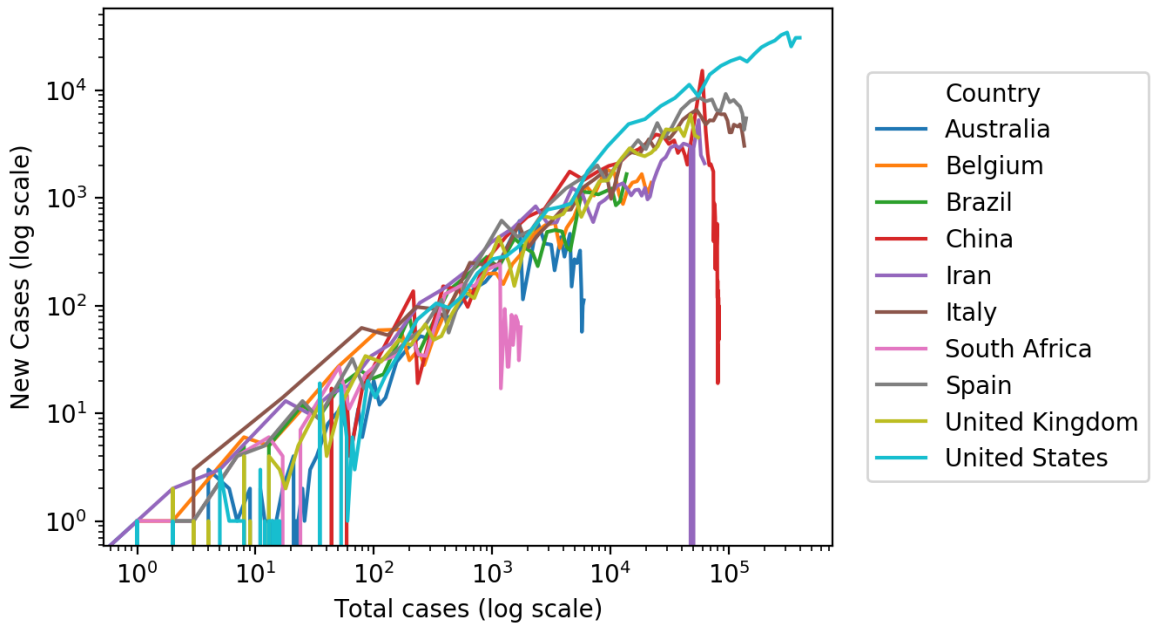

Figure 02: New daily cases by the total number of cases, in logarithmic scale.

Here we propose that an alternative way of representing the dynamics of the virus in a population is through a normalized curve of identified cases. In such a graph (shown in Figure 3) the number of cases is shown in the Y axis as a proportion of the number of tests (for example, number of positive cases by 100 tests). X axis represents a given time window (days in Figure 3). This type of graph eliminates the major caveat of the type shown in Figure 1, with absolute number of cases, which does not take into account the number of tests performed in a given period. There are, however, several caveats with the proposed visualization format shown in Figure 3. First, data reporting the number of tests per day performed in a given country is more difficult to obtain. This precludes the format of being used widely for several countries. Second, the criteria for selecting who will be tested vary significantly between countries or even within the same country in different periods of time. Since most countries are not testing a random set of samples of their populations, the numbers we have available are coming mainly from symptomatic patients who are tested for the presence of the virus. This is an important limitation and certainly affects the interpretation of the data. Finally, the number of tested people also depends on test availability. Not all countries have tests at hand to everybody. Therefore, the analyses can be misleading even if the proportion of the number of tests is used.

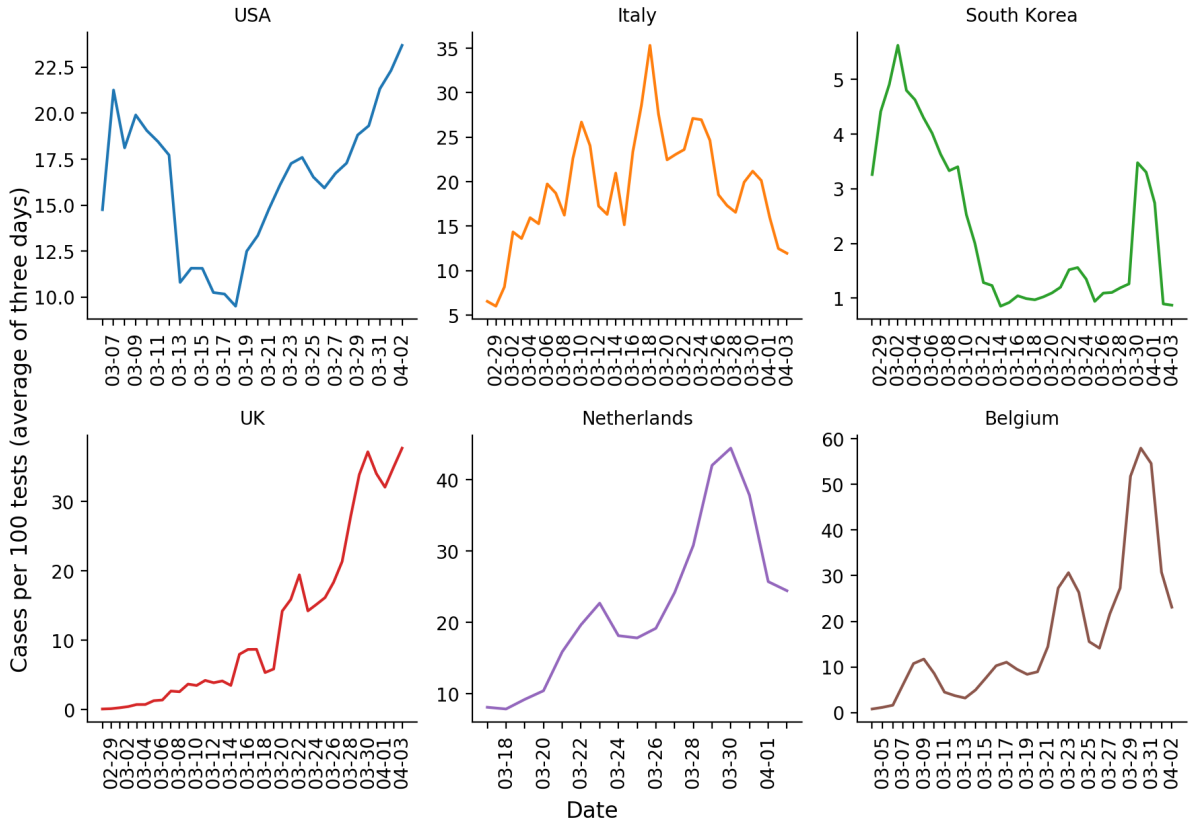

Figure 03: Normalized cases (cases per 100 tests) for six countries from February 15th to April 6th.

Nevertheless, the analysis of the relative number of cases in five countries allow some insights on the population dynamics of the virus. Take Italy as an example. The curve shown in Figure 3 suggests that a peak was reached around March 18 and since then the relative number of cases has decreased significantly. United States is another interesting example. Figure 3 shows that a first peak was reached around March 7 with a reduction of relative cases in the next 10 days' period with a subsequent rise after that. This is expected for countries with continental dimensions where the virus dynamics will behave like multiple waves each one based on a region of the country. The first cases in the United States were reported in the west coast with a subsequent peak in the east coast, first in the state of New York. These different waves of spreading are better displayed and interpreted with a type of visualization seen in Figure 3.

Another interesting feature that can be used to compare different countries is to avoid calendar dates. Instead of using such dates, like February 1st, on the X axis, it is better to show how the curves behaved during the days after the N^{th} reported case. For example, the country curves can be compared after the 10th patient was diagnosed, regardless of the calendar date on which it occurred.

Finally, one important type of data is the number of deaths, since those are absolute numbers affected by a smaller number of variables. In Figure 4 we present the number of new deaths (Y axis) compared to the total cumulative number of deaths (both in logarithmic scale). This way of representing the data seems to reliably reflect the virus population dynamics. Take, for example, the curve shown for China. After a peak in middle of February, the rate of new deaths drops to almost zero in the subsequent weeks. On the other hand, USA and UK seem to be going through an exponential growth of deaths, according to Figure 4.

The way information is transmitted to the general public directly impacts both the level of knowledge about COVID-19 and the epidemic itself, as well as people's attitudes to this crisis [10]. As shown here, different data visualization formats are more effective in providing reliable information to the general public and may help to make complex data more accessible, understandable and usable.

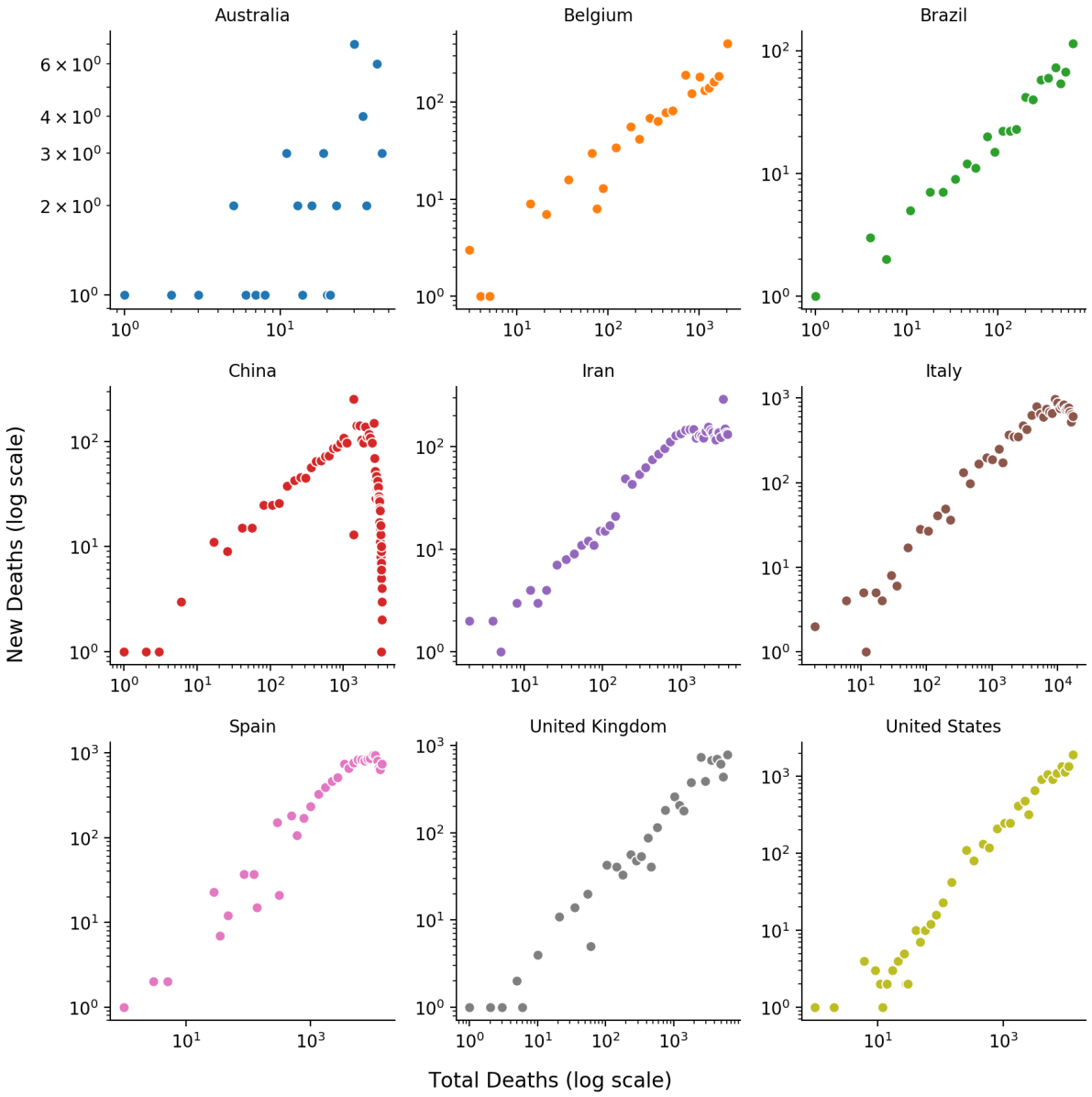

Figure 04: Daily new deaths by the total number of deaths, in logarithmic scale.

To make this and other data available to both the academic and general public, we have developed a specific web portal: https://bioinfo.imd.ufrn.br/covid-19. We will continue to monitor the COVID-19 pandemic in the next few months using the strategies and data formats discussed here. We envisage that our portal will evolve to a general platform involving data from outbreaks involving other pathogenic agents.

Competing interests

The authors declare no competing interests.

References

1- COVID-19 Statistics Policy Modelling and Epidemiology Collective. Defining high-value information for COVID-19 decision-making. 2020. medRxiv: 10.1101/2020.04.06.20052506.

2- Diaz-Quijano, FA; da Silva, JMN; Ganem, F; Oliveira, S; Vesga-Varela, AL & Croda, J. 2020. medRxiv: 10.1101/2020.04.05.20047944.

3- Akhtar, IH. 2020. Understanding the CoVID-19 pandemic curve through statistical approach. medRxiv: 10.1101/2020.04.06.20055426.

4- Dey, SK; Rahman, M; Siddiqi, UR & Howlader, A. 2020. Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. Journal Med. Virology 2020: 1-7.

5- Ganem, F; Mendes, FM; de Oliveira, SB; Porto, VBG; de Araújo, WNA; Nakaya, HI; Diaz-Quijano, F & Croda J. 2020. The impact of early social distancing at COVID-19 outbreak in the largest metropolitan area of Brazil. medRxiv: 10.1101/2020.04.06.20055103.

6- Roser, M & Ritchie, H. 2020. Coronavirus Disease (COVID-19) – the data. Published online at OurWorldInData.org.

7- H. Wickham. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag: New York.

8- Hunter, JD. 2007. Matplotlib: A 2D Graphics Environment. Comp. Science Engineering 9: 90-95.

9- Valdiserri, RO & Sullivan, PS. 2018. Data visualization promotes sound public health practice: The AIDSvu example. AIDS Educ. Prev. 30:26-34.

10- Qazi, A.; Qazi, J; Naseer, K; Zeeshan, M; Hardaker, G; Maitama, JZ & Haruna, K. 2020. Analyzing situational awareness through public opinion to predict adoption of social distancing amid pandemic COVID-19. J. Med. Virol. 10.1002/jmv.25840.