Building a statistical surveillance dashboard for covid-19 infection worldwide

Stefano Barone[1], Alexander Chakhunashvili[2], Albert Comelli[3]

[1]University of Palermo, Italy; [2]Karolinska University Hospital, Sweden [3]Ri.MED Foundation, Italy

Abstract

Background

When a pandemic like the current novel coronavirus (covid-19) breaks out, it is important that authorities, healthcare organizations and official decision makers, have in place an effective monitoring system to promptly analyse data, create new insights into problematic areas and generate actionable knowledge for fact-based decision making.

Aim

The aim of this article is to describe an initial work focused on building a comprehensive statistical surveillance dashboard for the epidemic of covid-19, which can be exploited also for future needs.

Methods

We propose novel ways of exploring, analysing and presenting data, using metrics that have not been used previously. We also show the steps necessary to build and operate such a dashboard.

Results

As a result of this this work, a set of data exploration and data visualization tools have been proposed which can be instrumental in providing information necessary to manage a crisis like covid-19 pandemic in a more systematic and effective way.

Conclusion

The proposed statistical surveillance dashboard can provide formal authorities and other decision makers with valuable insights into problematic areas and help them make critical decisions based on facts and an in-depth data analysis. The dashboard is implemented in an online dedicated website, freely accessible to the readership of this journal.

Introduction

A novel coronavirus started its spreading in Wuhan, China in December 2019, causing severe disease in human beings in most of the cases and death in a limited percentage [1], [2], [3]. Since the initial outbreak the epidemic of the virus (hereinafter named covid-19) has rapidly spread to cover almost all countries and territories of the world. More than two million people are already infected by covid-19 and many more are expected to get the virus until a remedy will be found and the virus spreading stopped or at least slowed down substantially. The risks for a spreading of the virus were immediately communicated by scientists to the community, also in Europe (see e.g. [4], [5], [6]). Despite the alerts, covid-19 hit Europe and western countries, spreading at high velocity (see e.g. [7], [8]).

In such a scenario, it is of utmost importance to set up an effective surveillance system that can promptly and correctly analyse source data, create new insights and supply actionable knowledge for fact-based decision making.

Several public institutions have set up websites for the visualization of the trends and free downloading of data (e.g. [9], [10]). The spatial visualisation of the spreading of the virus is important. In [11] the authors analyse the spatial representation of the disease, material, population and social psychology at three scales: individual, group and regional scale. At the individual scale, the comparison between spatial epidemic tracking and the spatiotemporal trajectories of patients was carried out. At the group scale, the estimation of population flows and the spatial distribution was carried out. At the regional scale, the segmentation of spatial risk, the analysis of balance between the supply and demand of medical resources, and the spatial differentiation analysis of material transportation capacity and social sentiment were carried out. Moreover, an understanding of how the contagion can happen within human population is incredibly important, see e.g. a study made in [12] where the authors have established some correlations between virus spreading and transportation means.

Based on observed data there is a streamline of research devoted to model the epidemic trends, by means of stochastic models mostly grounded on mathematics (see e.g. [13]). These models tend to mainly estimate the proceeding of the epidemic.

However, all these models contrast with three aspects:

1. Novelty of the virus, which has possibly new ways of transmission and evolution;
2. Containment policies make the virus spreading not free, so the models have to duly take these aspects into account; policies may change day by day, from place to place.
3. Stochastic-mathematical models need time to be developed and calibrated, while we need very immediate tools for surveillance and decision making.

Therefore, we are oriented to a more descriptive, exploratory data analysis, which requires zero or very few relatively weak assumptions. Such an analysis is also more versatile and faster to perform. Moreover, an exploratory data analysis method is easier to be understood by decision makers and general public. A first step in this direction was proposed by [14] for the analysis of covid-19 data.

Methods

Data source

We use the dataset daily published on the site of European Centre for Disease Prevention and Control [9]. We assume that data are fair or at least they have been produced using the same formal methods, although we are aware that there may be computational differences from country to country, due to different standards, different data handling procedures and other deviations that can influence this type of activity in emergency situations. Problems with multisource data gathering were already well posed and described by [7].

Dataset pre-processing

The daily dataset is pre-processed using a code developed in MATLAB©, that has the specific function of reordering data according to the country of origin, creating an array where data are reported by columns “reporting date”, “number of daily cases”, “number of daily deaths”, “country”, “population size” etc.

For each country, the number of daily cases reported is examined and all cells reporting zero cases prior to the first positive case are eliminated. This operation is functional to start the monitoring of the epidemiological process in each country as from the moment when the first positive case was found. This is

a particularly demanding assumption. Even though there have been cases before, they were not reported for various reasons, on which we do not comment but we are aware that it is a particularly important and delicate aspect. Cells containing the zeros preceding the first non-zero value are eliminated along with the other corresponding cells of the other two columns of the same country.

After the pre-processing, the array is saved as a worksheet, ready to be further analysed.

Dataset analysis

The cleaning and statistical analysis of the dataset is currently done with Microsoft Excel© which is a widespread tool for data analysis. A first validity check of the dataset is made, to examine if there are correct data for each originating country.

The countries that have supplied data for at least 20 days and only columns reporting date, number of cases, country of origin are currently analysed.

The analysis is accordingly focused on countries where there are enough data to conduct a statistical analysis. The dataset is then processed and for each country the following indicators are calculated:

T: number of days from the beginning of the epidemic process, i.e. time from the first reported infection case.

N: total number of reported infection cases. It is calculated as the sum of the daily reported cases. it represents a measure of the spreading of the epidemic process in each country.

N/T: for each country the ratio N/T is calculated. This ratio is the average of daily cases calculated over the entire time span from the epidemic start and can be seen as an *average speed of the epidemic process* spreading, in analogy with physics. If the total number N of infection cases is seen as the distance travelled and the number of days T seen as the time passed from the motion start, the ratio N/T provides the average speed of the epidemic motion.

The list of countries is then rearranged by decreasing N/T value. Thereafter, a table is constructed in such a way that the countries with the highest N/T value are placed at the top of it, while the countries with the lowest N/T at the bottom.

As a result, the table shows which are the most critical situations worldwide that need more attention and require more synergy from all countries. The table can also prompt comparisons over time, to check possible improvements of worsening of conditions for each country. The table can be a good summary for decision makers to see “where are we?”, “are we improving or worsening?”, “are we in a good condition?” and if yes “how to keep such a good condition?” If not “how could we act to improve?”

Furthermore, the table shows which are the countries needing more attention, where the epidemic process is in very early stages and therefore can be much more easily stopped. Countries being in the intermediate zone can take all precautions to ensure that the epidemic process does not accelerate.

Plotting the N/T ratio by time is a valid help to monitor the progress of the epidemic. This type of graph and its special utility will be seen in the Results section.

For the most critical cases in relation to the aforementioned N/T ratio, graphs are drawn up, showing the time series – so the trends – of the daily cases and the cumulated cases. These series start from the starting date of the epidemic process, as previously specified. The cumulated cases are obtained by cumulating day by day the number of new cases observed. These graphs provide a ready-to-view picture to understand how epidemic processes are evolving in the most critical situations. Therefore, they help understand if and how trends are growing, getting slower or stabilizing. Moreover, they can be used to immediately compare the

absolute values of total cases between countries in order to make relevant/effective decisions of moving energies from one place to another.

A plot of the cumulated cases by date for the most critical countries provides a summary framework to understand what has been the sequence of development of the epidemic processes in the countries. On the other hand, this type of analysis, although provides valuable information for future epidemic cases, it does not have great value for the purposes of comparison.

It is even more interesting to compare the curves of the cumulated cases by replacing the actual reporting date with the T, for each country, starting from the start date of the epidemic process. In this case the curves in the monitored countries are represented as if the epidemic processes were starting at the same time.

In addition, the plots of the daily cases can help monitoring the situation day-by-day, to help decision makers understand if there are special issues to be addressed, and to check if the reporting process has validity, as the presence of outliers could be symptomatic of reporting errors.

However, more interesting are the plots obtained by calculating the differences of "new cases" between two consecutive days. These points show the daily jumps (increases/decreases) of infection cases. Keeping the anticipated analogy with physics, this plot can be seen as the sequence of accelerations/decelerations of the epidemic motion.

In general, we expect that after a first phase of acceleration of the epidemic process (hence positive values on the plot) a "braking" phase follows, in which the acceleration is negative (deceleration). After those two phases, we expect a stabilization phase in which the acceleration keeps around the zero value. Obviously, this is a theoretical reasoning, as the daily reported data are affected by unavoidable random variation.

For the stabilization phase, and to have another comparison perspective, we propose another indicator

$\boldsymbol{N'/T'} = \frac{\boldsymbol{N/N^*}}{\boldsymbol{T/T^*}}$, where N* is the population size of each country and T* is a reference time, which can be initially taken the same for all countries. So formulated, N'/T' is a dimensionless indicator which can also be used for comparisons between countries.

Results

Table 1 reports the summary table of the analysis of data downloaded on April 17th 2020. The table shows for each country the numbers of cumulated infection cases (N), the number of days from the epidemic start (T) calculated according to the method described above, and the N/ T ratio. The countries are sorted from the highest to lowest N/T ratio. Three additional columns report the N/T ratio calculated previously and the column "condition progress", which can be worsening or improving depending on the progress of N/T. in the table, the condition progress is reported only for the countries of the euro-zone.

Figure 1 shows the plots of the N/T ratio by epidemic time span, i.e. days from the epidemic start in the respective country. The figure presents four illustrative cases, the first two refer to China and South Korea. From those examples it can be seen that the average speed of the epidemic can be slowed down substantially in a rather short timeframe. This is the type of trend which is desirable to see in any country. The third case (c) shows the N/T ratio in Italy. It can be seen that the average speed has been reduced, but it's still growing after sixty days. The fourth plot (d) is at regional level, comparing the trends of the N/T ratio within Sweden and showing that Stockholm region is evidently more critical than others.

Table 1 and Figure 1 are complementary. The former is more valuable for comparison between countries at a specific moment as it provides a snapshot of the current situation. The latter is more valuable for comparison within countries and for comparison in terms of dynamics of the epidemic phenomenon.

Table 1. Summary table reporting number of cumulated cases, time lag from epidemic start, and its ratio. Updated 17th April 2020.

COUNTRY	N	T	N/T (17/04/2020)	N/T (09/04/2020)	N/T (03/04/2020)	Condition progress
United_States_of_America	671331	88	7628,8			
Spain	182816	77	2374,2	2125,9	1749,8	WORSENING
Italy	168941	78	2165,9	1991,7	1800,7	WORSENING
Turkey	74193	37	2005,2			
Germany	133830	81	1652,2	1482,2	1097,3	WORSENING
Iran	77995	58	1344,7			
United_Kingdom	103093	78	1321,7			
France	108847	84	1295,8	1079,6	844,4	WORSENING
China	83754	109	768,4			
Brazil	30425	52	585,1			
Netherlands	29214	50	584,3	489,3	408,3	WORSENING
Switzerland	26651	52	512,5			
Belgium	34809	74	470,4	354,6	255,8	WORSENING
Portugal	18841	46	409,6	9,1	7,2	WORSENING
Russia	27938	77	362,8			
…	…	…	…	…	…	…
Montserrat	11	28	0,4			
CuraÃ§ao	14	36	0,4			
Greenland	11	29	0,4			
Central_African_Republic	12	33	0,4			
Saint_Vincent_and_the_Grenadines	12	36	0,3			
Seychelles	11	34	0,3			
Suriname	10	34	0,3			
Gambia	9	31	0,3			
Papua_New_Guinea	7	28	0,3			
Mauritania	7	34	0,2			
Nepal	16	84	0,2			
Holy_See	8	42	0,2			
British_Virgin_Islands	3	22	0,1			
Anguilla	3	22	0,1			
Bhutan	5	43	0,1			

Figure 2 shows the plots of the cumulated cases (orange colour) and the daily new cases (blue colour) for the six most critical situations in descending order of criticality on date 17 April 2020. As already mentioned the criticality is based on the N/T ratio.

Figure 3 shows the curves of the cumulated cases, by respective reporting date for the nine most critical countries as of the 12th of March 2020. At that date it was visible a big gap between China and the other countries which were hit later by the epidemic. It was also visible the behaviour of the epidemic process in South Korea, which was already stabilising.

This type of graphical analysis allows an effective comparison between the evolution of epidemic processes in the currently critical countries. It also highlights the sequence of epidemic starts, so giving an idea of how the virus has advanced geographically.

These plots show if countries manage to promptly and effectively react to the epidemic with strong containment policies. Already on March 12 we could see that China and South Korea were able to flatten the curve, by abating the daily new cases to relatively low numbers.

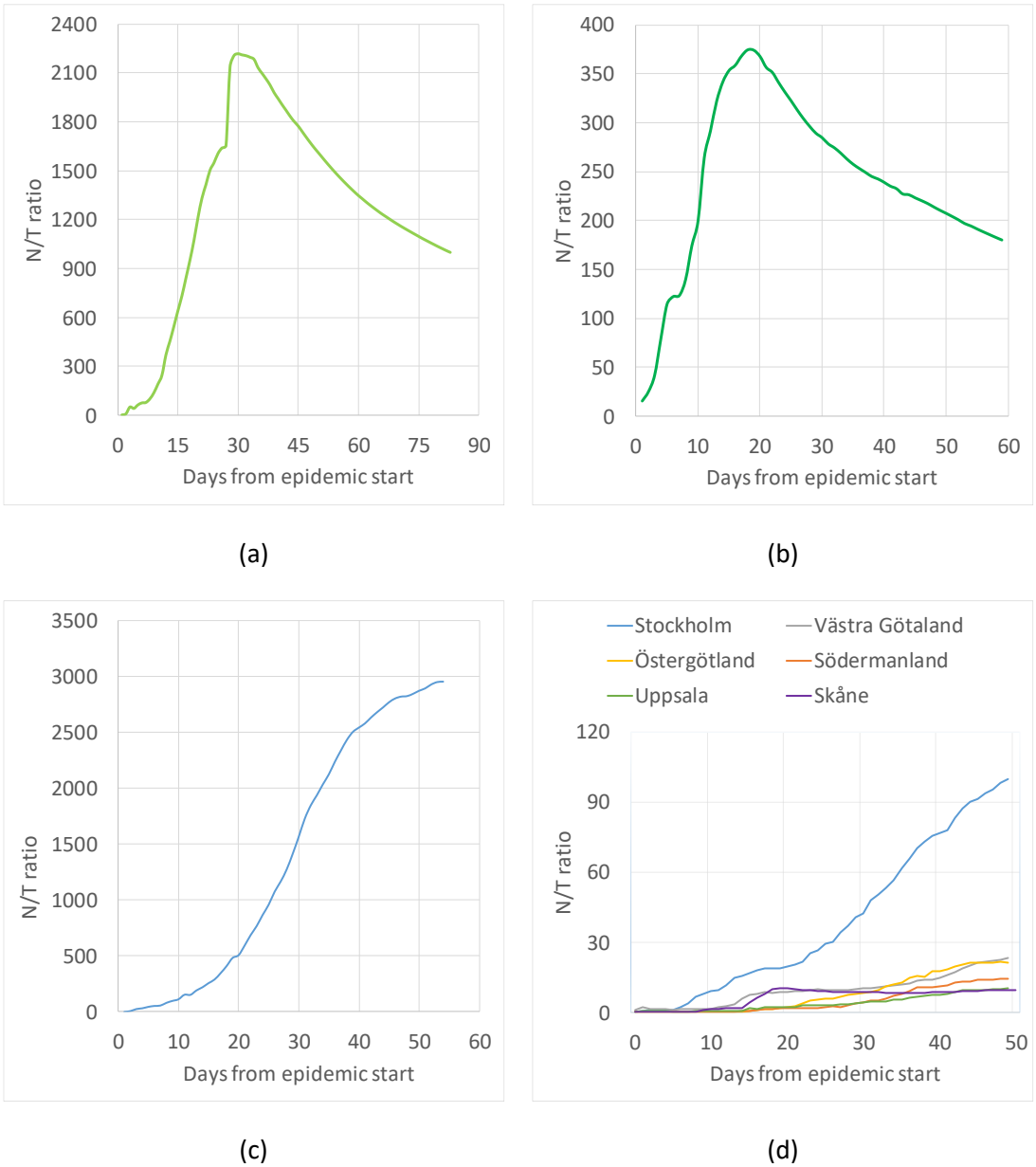


Figure 1. Plot of the N/T ratio by time from the epidemic start in (a) China, (b) South Korea, (c) Italy, and (d) six regions of Sweden (updated 15th April).

Figure 4 shows the plots of the cumulated cases by time from the epidemic start of the five most critical countries on 15th April 2020. These type of plot allows better comparison between trends of the epidemic process in different countries, showing for instance that USA is proceeding much faster than the other countries, Spain surpassed Italy, and that Germany is mostly aligned with Italy in terms of spreading of the virus. Iran is also critical, but running slower. The speed of infection spreading, which is in practice given by

the curve of the daily new cases, and in average given by the N/T ratio, can be due several concomitant reasons. Excluding the virus aggressiveness on human beings, which can be assumed the same all over the world, the difference of spreading speed could be due to mobility of people, which despite restrictions imposed by lockdown policies, is still able to generate occasions of contagions. Moreover, there might be differences between countries due to different policies to test people for the presence of virus. In countries where it has been opted for a mass screening, in the initial phases of the epidemic the number of daily new cases of infection tends to inflate. However, some countries have demonstrated the policy of mass screening can be then beneficial in terms of faster abatement of epidemic spreading, if other policies of containment are adopted accordingly (e.g. contact tracing).

Figure 5 (a) and (b) show the plot of daily increments/decrements for China and South Korea separately. The expected behaviour which should be initially mostly positive (epidemic acceleration), then mostly negative (epidemic controlled braking) then stabilizing around zero (controlled containment), although the unavoidable random variation, can be clearly seen in the data of China (a) and South Korea, which therefore represent model countries as they managed well the epidemic process spreading.

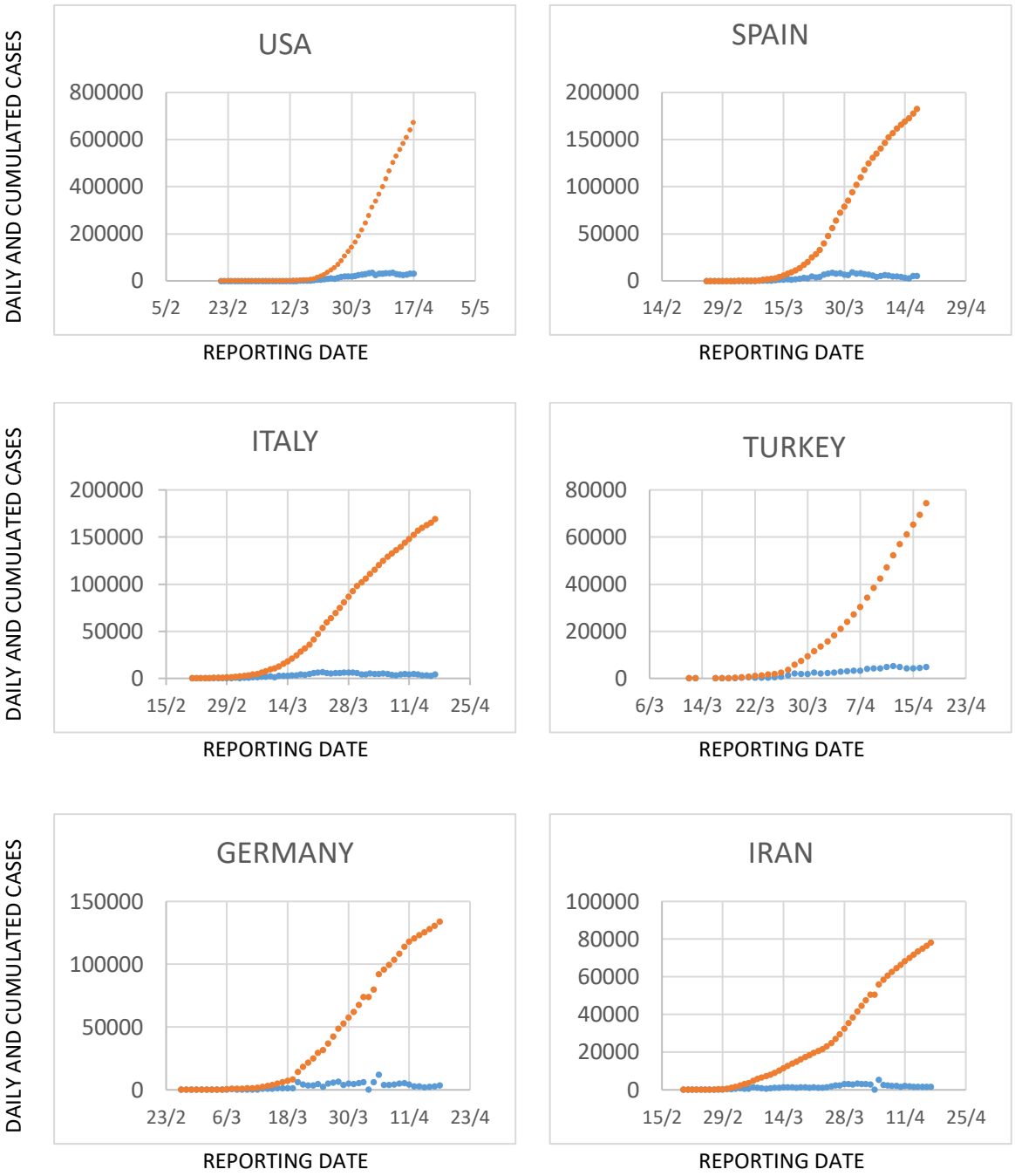


Figure 2. Plots of daily cases (blue dots) and cumulated cases (orange dots) by reporting date, for the six most critical countries (updated17/04/2020).

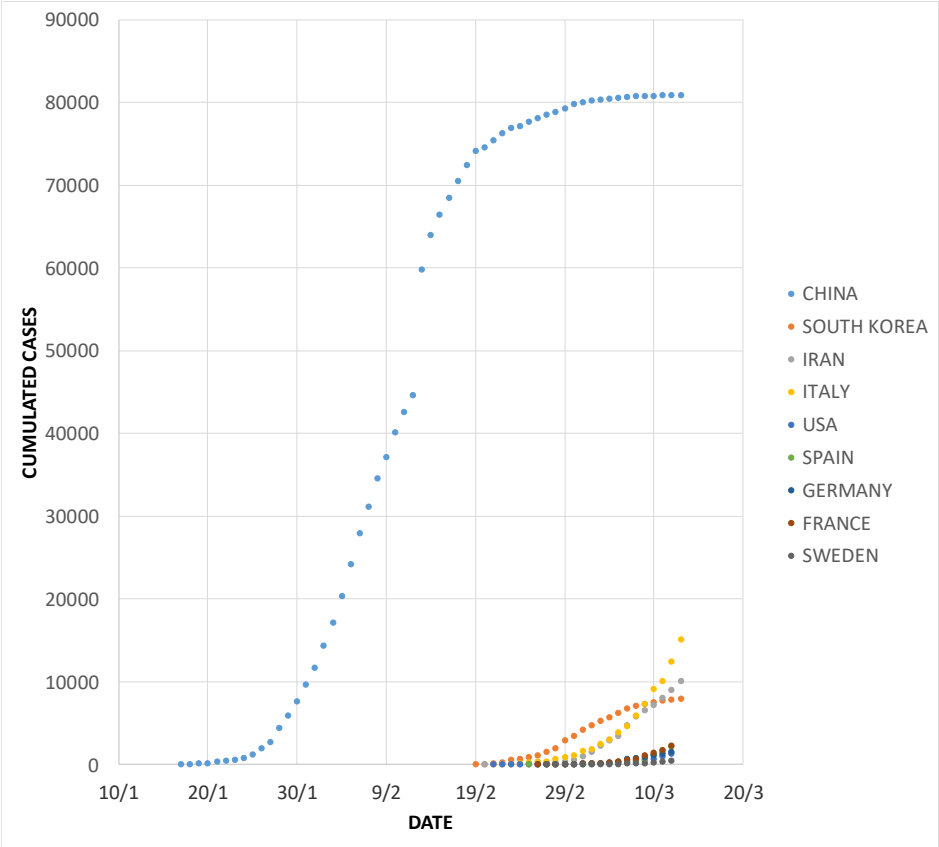


Figure 3. Cumulated cases by reporting date for most critical countries (updated 12 March 2020).

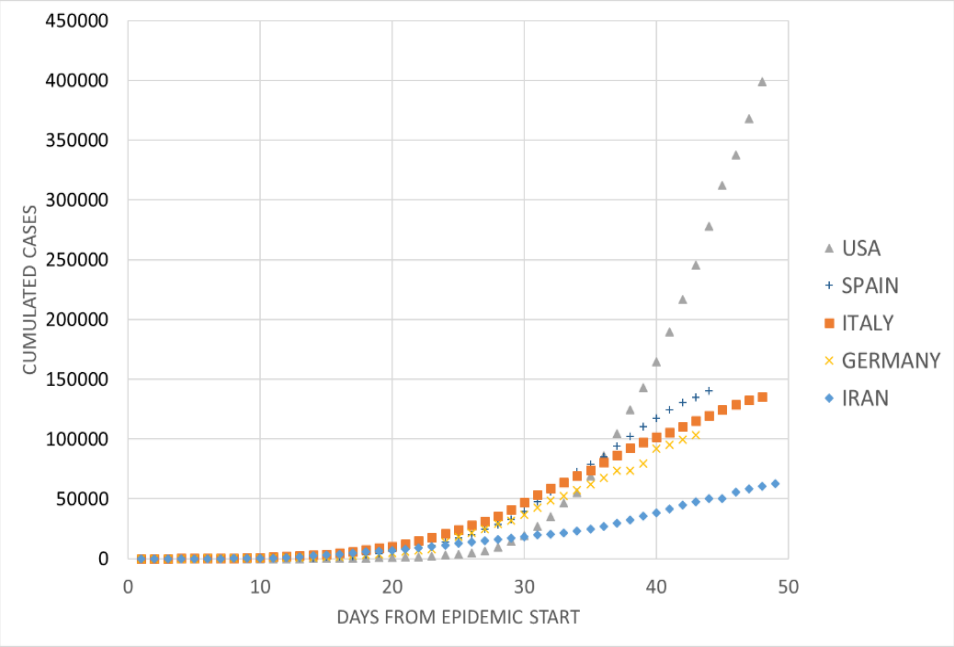
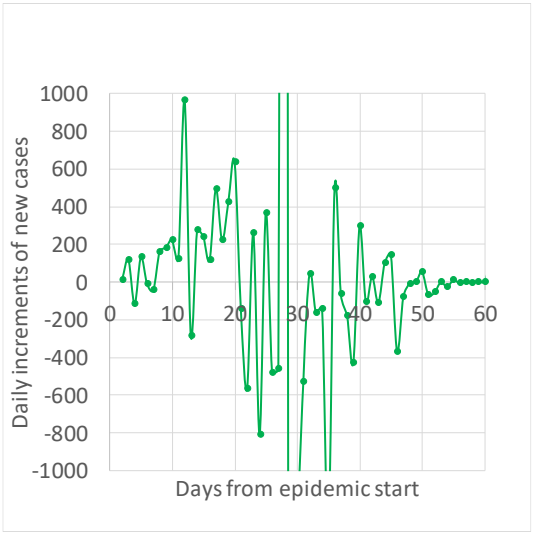


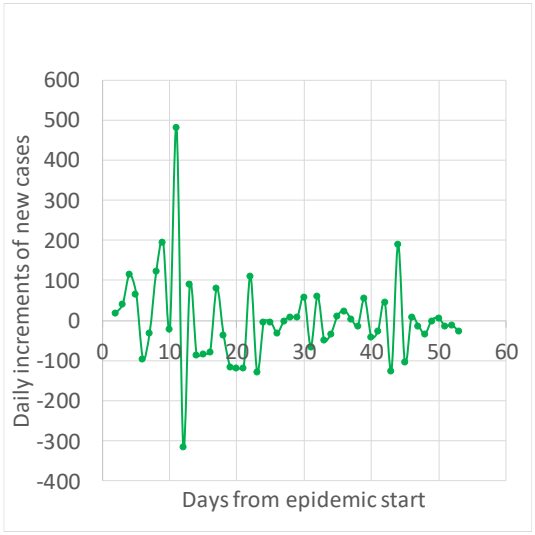
Figure 4. Cumulated infection cases by time from the epidemic start (updated 15/04/2020).

South Korea presents lower numbers than China, so we can take it as a benchmark for comparison with other countries. By way of example, we can see how the historical series of acceleration/deceleration of the epidemic process has evolved in Germany. It can be seen from the graph that the variation increases very rapidly, so we can say that the epidemic spreading process goes out of control (or containment). therefore, even if an average reduction in daily cases is achieved, variation is still extremely high and, therefore root causes of such high variation should be found and the relative issues solved. On the other side, we can see that the Scandinavia countries (Figure 5. (d), (e), (f)) are acting better to contain the epidemic especially Denmark and Norway, a bit less Sweden which seems to be going out of containment.

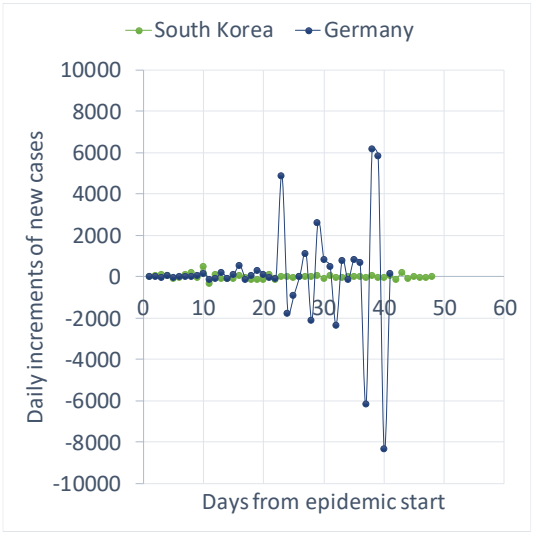
The South Korean case deserves to be deeply studied and understood. Some good insights were reported in [15] and [16].



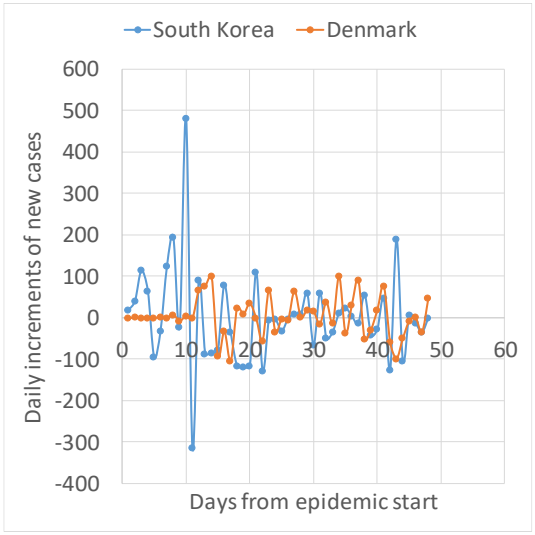
(a)



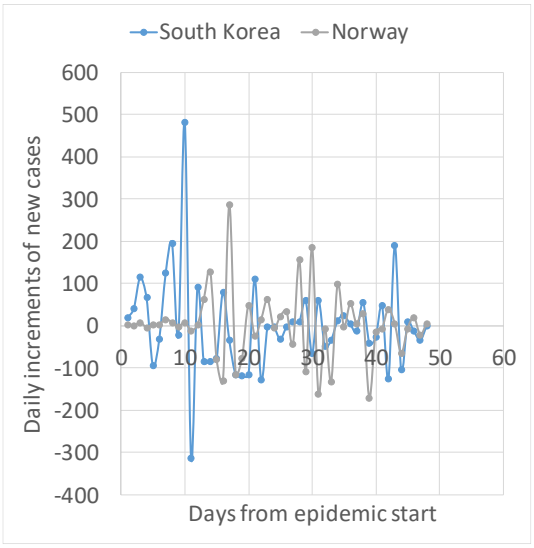
(b)



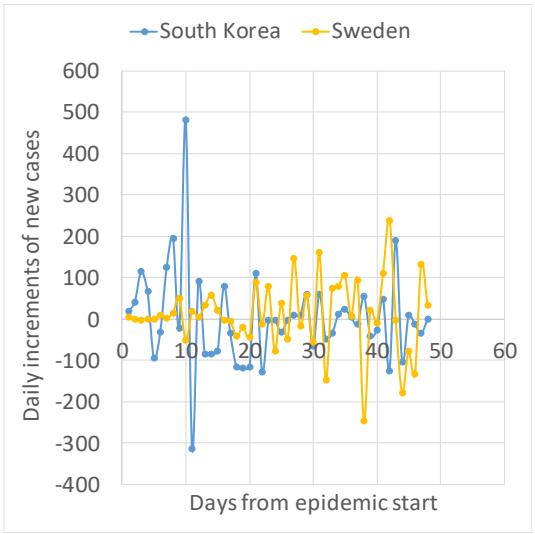
(c)



(d)



(e)



(f)

Figure 5. Increments/decrements of daily new cases (a) China, (b) South Korea (c) comparison Germany vs. South Korea, (d, e, f) comparison Scandinavia countries vs. South Korea.

Last but not the least, taking as reference time T* the least time to reach the situation of zero contagions, which is approximately 60 days for South Korea, the calculation of the dimensionless indicator N'/T' allows a comparison between countries with respect to the proportions of the epidemic compared to population sizes. Figure 6 shows the bar chart of the indicator N'/T' calculated for the Euro-zone countries and USA at five times during the most critical pandemic phase.

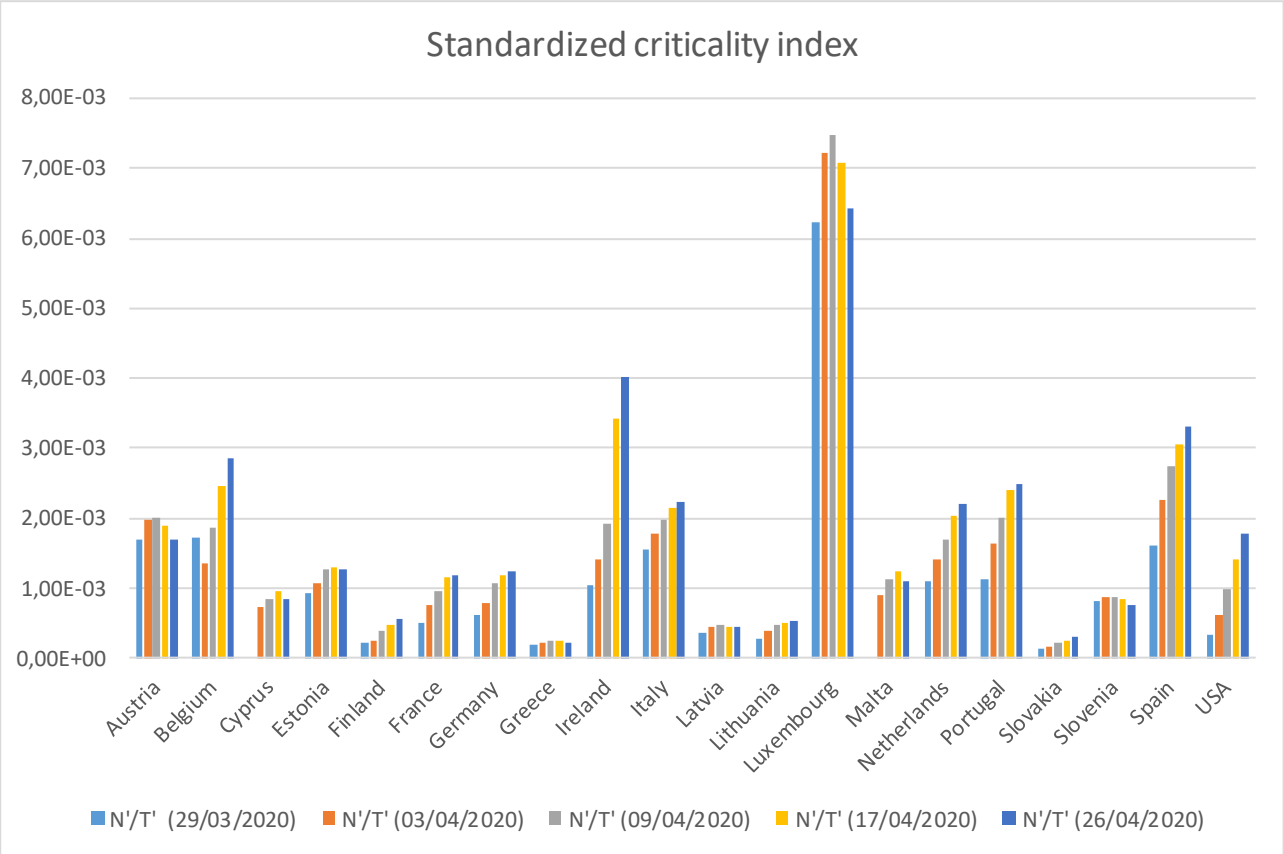


Figure 6. Dimensionless indicator N'/T' calculated for Euro-zone countries and USA at five times during the most critical epidemic phase (updated 26/04/2020).

Seen in this way, the perspective is very different. The worst regional scenario happens in Luxembourg, with big distance from the others, although it is improving. The case of USA doesn't look so dramatic, compared to most of the European countries. In the Euro-zone, there are situations worse than Italy, for instance, Belgium, Ireland, Netherlands, Portugal and Spain.

What to say about the comparison in terms of N/T or N'/T'? We can say that at the start of the epidemic, in the "fast burning phases" a comparison in absolute terms is more appropriate, because it is like treating all single lands as spots of fire and it is obvious that the more population density in a place the more likely to have big numbers of contagions in short time. However, with time passing, and with the expansion of the epidemic, it becomes better to see the things in relative terms, looking at the damages in proportion to the population sizes.

We can summarize by saying that numbers count, but proportions count, too.

Discussion

In this article we propose a set of novel tools and metrics for analysing covid-19 data in an explorative and non-inferential way (no predictive stochastic models have been employed). The daily collected data downloaded from the website of the European Centre for Disease Prevention and Control are thoroughly analysed. The performed analyses create a surveillance dashboard including graphs, figures and criticality tables useful for those who have to make political decisions on territories affected by the Covid-19 epidemic. The criticality table can be complemented with geo-demographical indications such as latitude and longitude of the country, entity of the population. These data can provide further details to better understand if there are more or less favourable conditions for the development of the epidemic process. These aspects will be the subject of a following study.

The presented analysis can be easily deployed within countries at regional level, and within regions at province levels. For example, one country can make a reasonable comparison between its different regions regarding the average velocity of the spread of the epidemic to identify the most critical areas requiring increased attention.

By using relatively simple explorative techniques to analyse data, we strive to provide authorities and other decision-makers with valuable insights into covid-19 epidemic, which can ultimately lead to vital improvements in both the containment strategy and the policies enforced across the countries or regions hit by the epidemic. The methods used are rigorous but not mathematically advanced making them easy to read and understand. Furthermore, it is done in a way that anyone interested can set up similar dashboards.

All the analyses presented in this article will be part of an online dashboard, where users will be able to select plots and tables of their own interest. The dashboard will be user friendly and accessible to the readership of this journal.

Acknowledgements

The authors thank their own institutions for the possibility to allocate time to do research on the topics addressed in the manuscript. The international collaboration between the authors is not new. Barone and Chakhunashvili did research together and produced several publications since 2003.

Conflict of interest

No conflict of interest of any kind is underlying the research work and the publication of the manuscript

Authors' contributions

All authors contributed equally to the definition of all steps of the research and in the writing of the manuscript.

References

[1] Velavan T.P. , Meyer C.G.. “The COVID-19 epidemic”. *Tropical Medicine and International Health*. volume 25 no 3 pp 278–280. march 2020

[2] Zhou P, Yang XL, Wang XG et al. “A pneumonia outbreak associated with a new coronavirus of probable bat origin”. *Nature* 2020. https://doi.org/10.1038/s41586-020-2012-7

[3] Andersen, K.G., Rambaut, A., Lipkin, W.I. et al. ”The proximal origin of SARS-CoV-2”. *Nat Med* (2020). https://doi.org/10.1038/s41591-020-0820-9

[4] Thompson RN. “Novel coronavirus outbreak in Wuhan, China, 2020: intense surveillance is vital for preventing sustained transmission in new locations”. *J Clin Med* 2020 Feb 11;9(2):498

[5] Michael P. Ward Xiangdong Li Kegong Tian. “Novel coronavirus 2019, an emerging public health emergency”. *Transboundary and Emerging Diseases*. DOI: 10.1111/tbed.13509

[6] Pullano G., Pinotti F., Valdano E., Boëlle P.Y., Poletto C., Colizza V. “Novel coronavirus (2019-nCoV) early-stage importation risk to Europe”, January 2020. Euro Surveill. 2020; 25(4). https://doi.org/10.2807/1560-7917.ES.2020.25.4.2000057

[7] Bernard Stoecklin Sibylle , Rolland Patrick , et al. ”First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures”, January 2020. *Euro Surveill*. 2020;25(6). https://doi.org/10.2807/1560-7917.

[8] Kinross Pete, Suetens Carl, et al. ”Rapidly increasing cumulative incidence of coronavirus disease (COVID-19) in the European Union/European Economic Area and the United Kingdom, 1 January to 15 March 2020”. *Euro Surveill*. 2020;25(11). https://doi.org/10.2807/1560-7917.ES.2020.25.11.2000285

[9] European Centre for Disease Prevention and Control (An agency of the European Union). Covid-19 data and dashboard.

[10] Coronavirus 2019-nCoV Global Cases by Johns Hopkins CSSE. (Available from: https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html

[11] C. Zhou, F. Su and T. Pei et al., “COVID-19: Challenges to GIS with Big Data”, *Geography and Sustainability*, https://doi.org/10.1016/j.geosus.2020.03.005

[12] Shi Zhao, Zian Zhuang, Jinjun Ran, Jiaer Lin, Guangpu Yang, Lin Yang, Daihai He, “The association between domestic train transportation and novel coronavirus (2019-nCoV) outbreak in China from 2019 to 2020: A data-driven correlational report” ,*Travel Medicine and Infectious Disease*, Volume 33,2020, https://doi.org/10.1016/j.tmaid.2020.101568.

[13] Zhao S, Lin Q, Ran J, et al. “Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak”. *Int J Infect Dis* 2020 (published online Jan 30)

[14] Dey SK, Rahman MM, Siddiqi UR, Howlader A. “Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach”. *J Med Virol*. 2020 Mar 3. doi: 10.1002/jmv.25743.

[15] HyunJung Kim. “South Korea learned its successful Covid-19 strategy from a previous coronavirus outbreak: MERS”., March 20, 2020. *Bullettin of the atomic scientists*. (https://thebulletin.org/2020/03/south-korea-learned-its-successful-covid-19-strategy-from-a-previous-coronavirus-outbreak-mers/)

[16] Ryu S, Chun BC, Korean Society of Epidemiology 2019-nCoV Task Force Team. “An interim review of the epidemiological characteristics of 2019 novel coronavirus”. Epidemiol Health 2020;42:e2020006