Geostatistical Investigations on the Spread of COVID-19

Földváry, L.
Department of Geodesy and Surveying, Budapest University of Technology and Economics, Műegyetem rkp. 3, 1111 Budapest, Hungary , E-mail: foldvary.lorant@epito.bme.hu

Abstract
A MATLAB tool has been developed for monitoring and analysing the spread of the COVID-19 virus. With the use of the tool geostatistical analysis of the continuously developing time series of contagion can be performed, which may be a useful tool for decisionmakers on planning future actions. The major features are geographical display of the status of the virus spread, time series of confirmed / fatal / recovered / active cases of the disease on cumulative and on daily basis, centre and radius of the spread, best fitting bell curve of the active cases, and general statistics (CRF, modified CFR, attack rate, mortality rate). Each feature can be used globally or country-wise. Results on 28 June 2020 indicates that the COVID-19 globally is in ascending phase yet, meanwhile in several countries worldwide a 2nd phase of contagion has been started.

1. Introduction

By the time this article worded, the COVID-19 disease has spread all around the world. The COVID-19 outbreak has been declared to be a pandemic by the World Health Organization (WHO) on 11 March 2020. There are already estimates on the severity of the virus based on the information from the case of China (Verity et al., 2020 and Wilder-Smith et al., 2020), though present status in Italy, Spain and the US indicate its severity rather high. Nevertheless, it has already turned out that it is a real challenge at to estimate reliably the transmissibility and the disease severity of COVID-19. A repeated RNA testing the entire 3000 residents of Vo'Euganeo, a completely isolated village in northern Italy has delivered unexpected results (Day, 2020). Anyone with positive test results was quarantined. The number of people with COVID-19 symptoms fall from 88 to 7, so by over 90 percent within 10 days. From the aspect of the transmissibility of COVID-19 it was observed that between 50-75% of the infected people, were asymptomatic, but contagious. We can say that we nearly know very few reliable information on the COVID-19. In consequence, opinion can be found on the media on the lack of rationale behind the decisions on eradication of the virus as they are unfounded in the lack of reliable data (Ioannidis, 2020). All this information indicate that the end of this pandemic is unpredictable by now, thus for the better understanding of the virus from all possible aspects is essential. The Geoinformatics can contribute to this issue by geographical monitoring of the spread and provide geostatistical analyses.

An interactive web-based dashboard is already hosted and operated by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, first shared publicly on January 22 (Dong et al., 2020). The CSSE dashboard displays the location and number of confirmed/fatal /recovered COVID-19 cases for all affected countries. It enables display of time series of infections globally and country-wise both on linear and logarithmic scales. This dashboard is so user-friendly that it became widely used by the public and more essentially by decisionmakers in determining actions to the evolving situation.

2. Data and Software

In case of a geostatistical analysis, a major issue is the reliability of the available data. The primary data source of CSSE dashboard is DXY, which is an online platform of the Chinese medical community (Dong et al., 2020). The information is then confirmed by them using regional and local health departments, namely the China CDC (CCDC), Hong Kong Department of Health, Macau Government, Taiwan CDC, European CDC (ECDC), the World Health Organization (WHO), as well as city and state level health authorities. All the data collected by CSSE team is made freely available in a GitHub repository, which is used as data source of the present analysis. Geostatistical analyses of the present study make use of this data without any critical overview its content.

For quantifying the extent of the spread of COVID-19, it is compared to the "spread" of the potential target group of the disease, that is the entire population of the Earth. Total population of the Earth is modelled using the "countries in the world by population (2020)" table of Worldometer

(Worldometer, 2020). The investigations were performed with MATLAB version R2018b. In MATLAB an analysis tool has been developed, in which one can manipulate, analyse, manage and provide map-based display of the COVID-19 data, also statistics and forecasting can be implemented. Nevertheless, though the representation is quite graphical, it is not a GIS as it has no interactive map-based graphical interface, queries cannot be performed on spatial basis. Some of the features are unique, which is to be introduced from mathematical aspect in the next chapter.

3. Methodology

The COVID-19 has gradually become from local through regional to a global phenomenon, and accordingly declared to be a pandemic. The geographical spread of COVID-19 can be measured by the taking into account in a way all geographical locations, where it has been confirmed. Obviously, those locations, where the number of confirmed cases is higher, the relevance of the location should be emphasized, i.e. such locations should be weighted. Based on statistical considerations, in a planar coordinate system the centre of the disease can be optimally described by the weighted centroid, i.e. weighted mean of the locations,

$$x_0 = \frac{\sum px}{\sum p}$$
$$y_0 = \frac{\sum py}{\sum p}$$

Equation 1

while the spread can be described by the weighted mean squared coordinate differences from the centre, and the radius of the spread is then determined from the coordinate components, i.e.

$$\sigma_x^2 = \frac{\sum p(x - x_0)^2}{\sum p}$$
$$\sigma_y^2 = \frac{\sum p(y - y_0)^2}{\sum p}$$

Equation 2

$$\sigma_r = \sqrt{\sigma_x^2 + \sigma_y^2}$$

Equation 3

The weights, *p* in Equation 1-3 are the confirmed cases of the virus, while *x* and *y* are rectangular coordinates. Such a measure may be valid for local (country) or regional (continental) scales, but cannot work globally, as for the Earth globally spherical coordinates are used unavoidably. Applying Equations 1-3 for a global spread may result false conclusion, as locations close to the 180° meridian affect the location of the centre randomly: even though the distance between Qamea Island and Kioa Island (both constitutes of Fiji Islands) is about 30 km, as Qamea Island is located at along the -179.8° meridian, confirmed cases from this island would shift the global centre to the East, while as Kioa Island can be found at the +179.8° meridian, it would push the global centre to the West. Obviously, proper solution can be achieved by using spherical coordinates and spherical distances between locations. Several solutions, similar to the present solution have been subjected by Buss et al., (2001), Panozzo et al., (2013) among others.

Accordingly, for defining the centre and radius of a disease, distances among the locations should be considered as spherical distance on the surface of the globe. The distance between a point A(φ_A λ_A) and the centre (φ_0 λ_0) is:

$$d = \text{acos}\,(\sin\varphi_A \sin\varphi_0 + \cos\varphi_A \cos\varphi_0 \cos\,(\lambda_A - \lambda_0))$$

Equation 4

and the centre is determined by minimization of the L2-norm of the weighted distance, that is:

$$\min\left(\frac{\sum pd^2}{\sum p}\right)$$

Equation 5

The minimization is performed using the Nelder-Mead (or simplex) optimization method (Nelder and Mead, 1965). As this method approximates a local optimum, for test purposes various initial coordinates have been defined resulting always the same location of the minimum, therefore the signal was found to be smooth. The derived centre and radius of the spread of COVID-19 are compared to that of the population of the Earth, serving as a reference. For this purpose, Equations 4 and 5 has been used for the whole population.

Another feature of the developed COVID-19 tool is displaying the future progress of the virus based on the available data. The assumption behind this prediction is that active cases of the virus follow normal distribution, which is theoretically the progress of a virus in a closed society with no interventions applied. Although this is obviously never be the case, still a fit of a Gaussian function (bell curve) can provide useful information, as deviations from a natural evolution of the virus spread indicates effects of implemented strategic actions, such as quarantines, isolations, hygiene regulations, suspensions and closures of institutions, or any other restrictions. Also, deviations from the normal distribution can observed due to appearances of the virus at different times resulting in waves of the spread interfering with each other. Nevertheless, the fit of a Gaussian curve is very uncertain at an early stage of the spread, but by time as the time series of the data gets longer, the estimates become more reliable and more informative. The equation of the Gaussian bell curve is:

$$y = a \cdot e^{-\left(\frac{x-b}{c}\right)^2}$$

Equation 6

In the present case the abscissa *x* refers to the date and the ordinate y is the active COVID-19 cases, that is confirmed cases minus the deaths and recovers. The estimation of its parameters, *a*, *b* and *c* is by non-linear least squares adjustment making use of the Trust Region restriction-step method (Moré and Sorensen, 1983). With the estimated parameters, the Gaussian bell curve is extrapolated mainly with the aim of displaying, but not considered to be a reliable forecast.

4. Status of COVID-19

The status of the COVID-19 disease can be described by 10145791 confirmed, 501893 fatal (death) and 5140899 recovered cases by 28 June 2020, which means 4502999 active cases. The temporal variation of the cases is shown on Figure 1. in logarithmic scale for the available days starting on 22 January 2020 until 28 June 2020. Figure 2 shows the new cases on daily basis, manifesting that the number of active and fatal cases increasing, they are on maximal value by these days. Figure 3 shows the present status of the confirmed cases for each country. Geographically distinct area of countries (e.g. Hawaii of US) are considered separately as from the aspect of the virus spread, they should be considered independent.

The dispersal of the virus is obviously found to be worldwide. The figure is more informative by zooming into the map for certain regions, countries, see the example for Central Asia on Figure 4. Beyond these displays, general statistics, spread metrics of the dispersal of the virus and best-fit bell curve parameters have been determined.

5. Spread Metrics of COVID-19

Based on these data sources, the spread metrics (spread centre and spread radius) has been determined using Equation 4 and 5 for the available days starting on 22 January 2020 until 28 June 2020. Figure 5 shows the dislocation of the spread centre by time with red asterisks. As a reference, the centre of the population is indicated by a blue asterisk. The centre obviously tends from the epicentre (Wuhan) to the West until May 2020. There is also a relevant move to the North is visible, this is, however, rather ostensible.

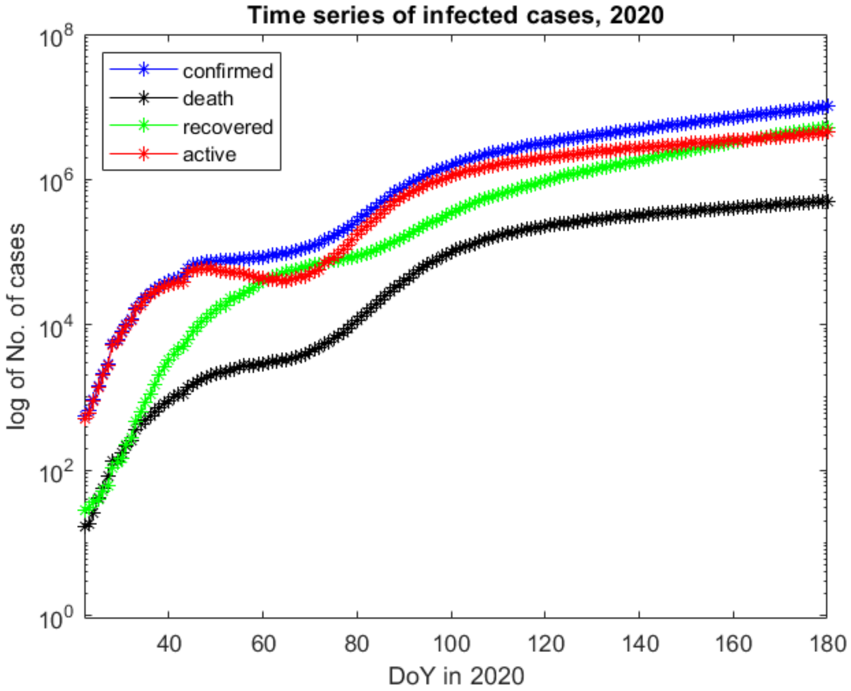

Figure 1: Time series of cumulative confirmed, fatal, recovered and active cases in logarithmic scale

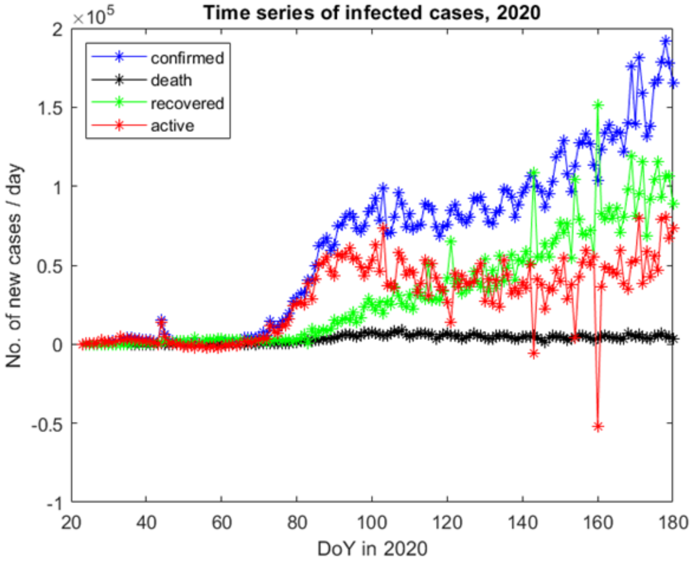

Figure 2: Time series of confirmed, fatal, recovered and active cases on daily basis

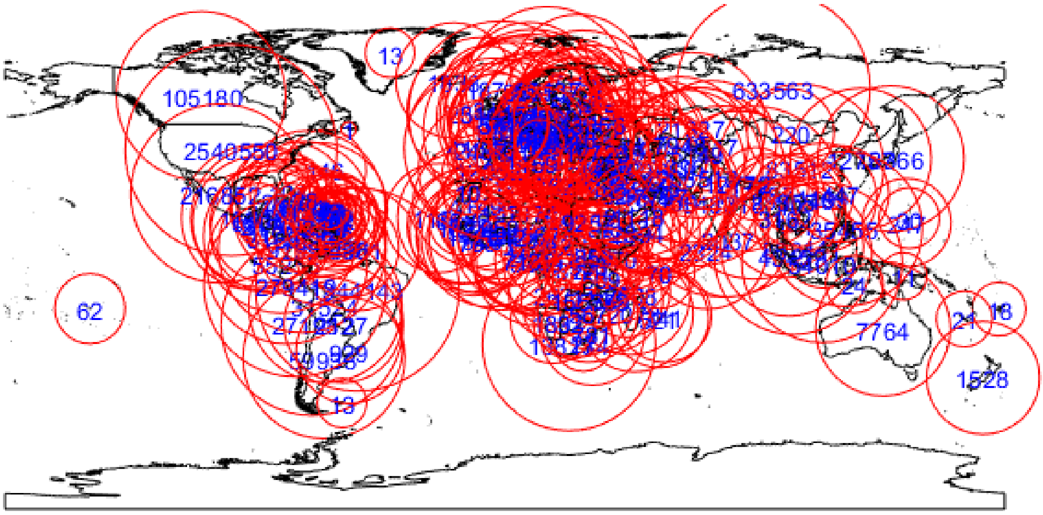

Figure 3: Display of confirmed cases for countries of the world

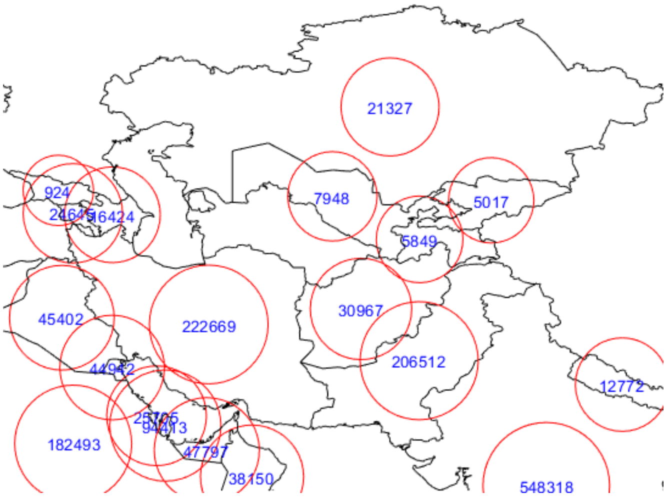

Figure 4: Display of confirmed cases by country

The fact is that the Mercator projection used for Figure 5 does not represent properly the globe nature of the Earth – this motion of the spread centre toward North is the effect of the emphasis of the Northern over the Southern hemisphere. By June 2020, it has turned to the Southeast. This is the consequence of the intense advancing of COVID-19 in Africa. This feature can more properly be illustrated in a polar projection of the globe, which also enables the displaying of the spread radiuses (Figure 6). The total population of the Earth in 2020 has a spread radius of 57.2311 degree (equivalent to 6371 km spherical arclength). By 28 June 2020, the spread radius is 61.5731 57.0122 degree (6854 km spherical arclength), which is more than the spread of the population; it is 107.6% of the latter. On the day, when the COVID-19 has declared to be a pandemic (11 March 2020), these values were 34.9355 degree (3889 km) and 61.50%.

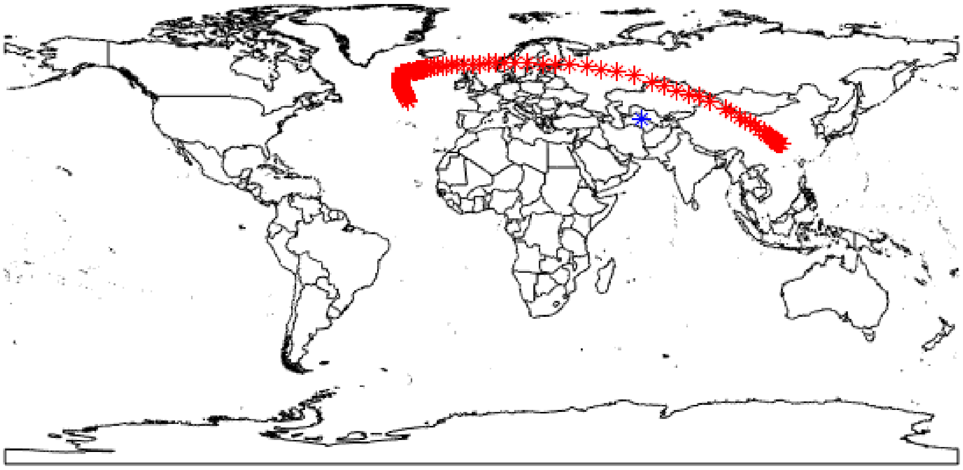

Figure 5: The location of the spread centre by time (red asterisks), and the centre of the population in 2020 (blue asterisk)

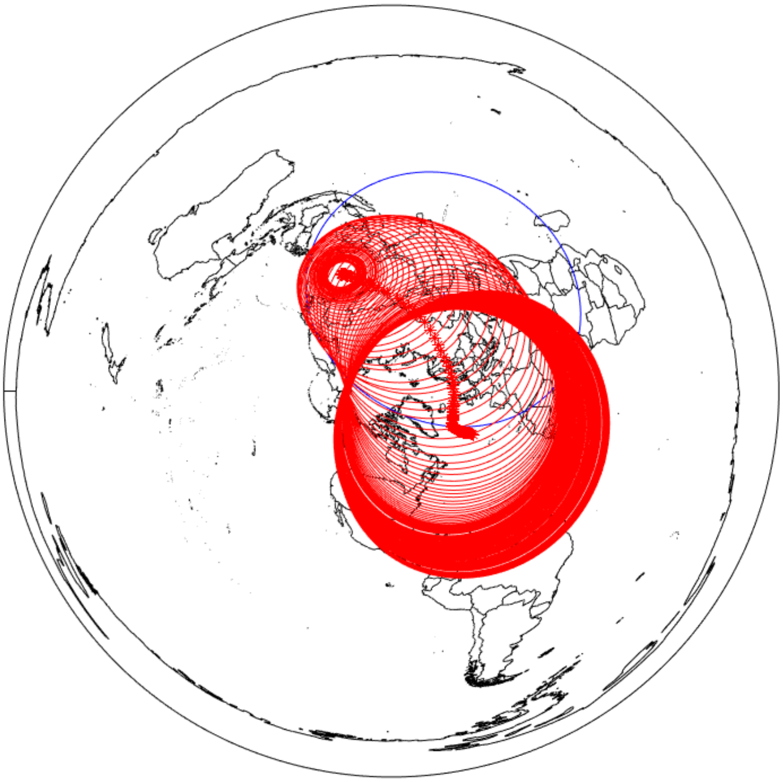

Figure 6: Visualization of the spread centre and spread radiuses by time (red curves), and the centre and the spread radius of the population in 2020 (blue curves)

6. Gaussian Distribution of Active Cases

Gaussian distribution can be fitted to the time series of the active cases for each country. Even though mathematically it can also be done for the global data, it is nonsense as Gaussian distribution can be expected only for groups of individuals, who meets the virus under the same conditions.

The example of Austria on 10 April 2020 is shown in Figures 7-10. Figure 8 shows that the active cases nicely follow the normal distribution, and that if everything would have remained the same, they would have overcome the virus by the end of April. This scenario could have been hold only if the conditions did not change. The drop of active cases can be nicely seen also in Figure 9, where also the major cause of the active cases can be identified with the recoveries. According to Figure 10, the daily number of confirmed cases also shows a decrease, which along with the decrease of the active cases indicates the retreating tendency. The status of Austria on 28 June 2020 are (displayed by the script): active cases: 551, confirmed cases: 17654, fatal cases: 702, recovered cases: 16401, CFR (case fatality rate, i.e. fatal/confirmed ratio): 3.98%, modified CFR (fatal/recovered+fatal ratio): 4.10%, attack rate (confirmed/population ratio): 0.1960%, mortality rate (fatal/population ratio): 0.0078%. As during the contagion period, CFR generally overestimates while modified CFR underestimates the actual CFR, the fatality rate can only be assumed as being within the two estimates. Therefore, in the case of Austria the fatality rate will probably remain in the 3.98-4.10% interval. Figure 11 shows the time series of active cases. It indicates that the estimates based on the Gaussian curve (Figure 8) turned to be too optimistic. Indeed, Austria has eased the restrictions in several ways in order to reduce the negative effects of the pandemic on industry and society as well. Consequently, the fading out of the virus is extended in time, which way the number of patients can be kept on a bearable level for the health care system.

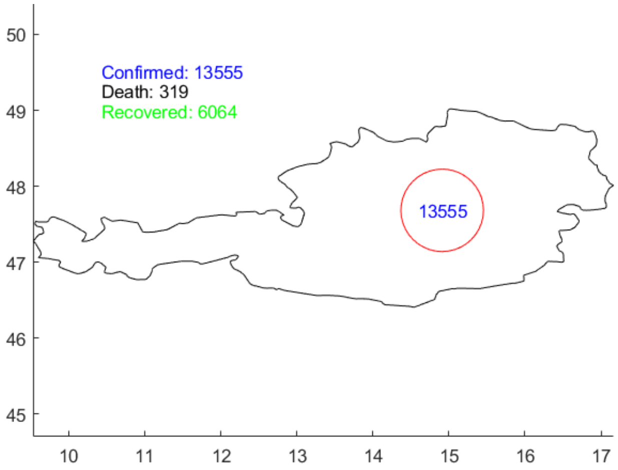

Figure 7: Visualization of the status of COVID-19 in Austria on 10 April 2020

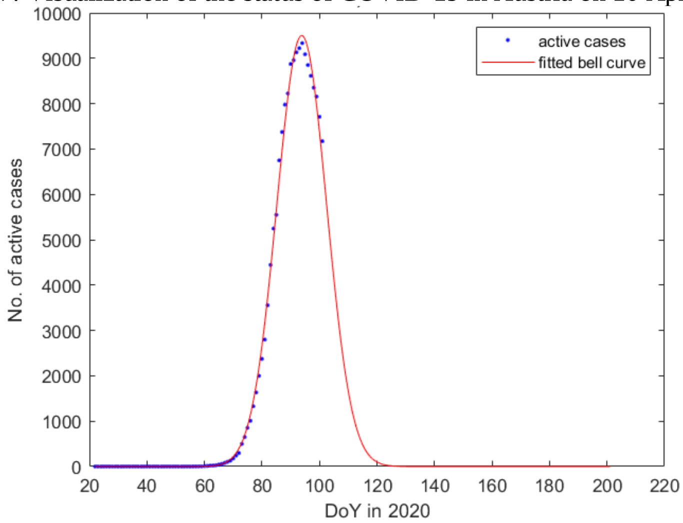

Figure 8: Estimated Gaussian bell curve based on active cases in Austria on 10 April 2020

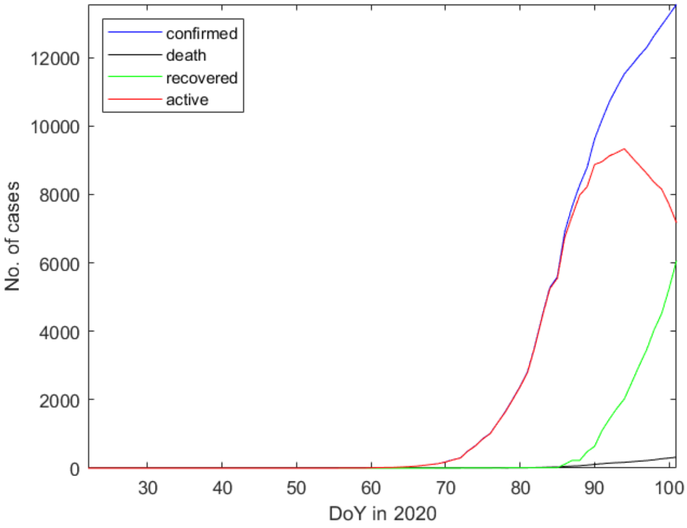

Figure 9: Time series of confirmed, fatal, recovered, and active cases in Austria until 10 April 2020

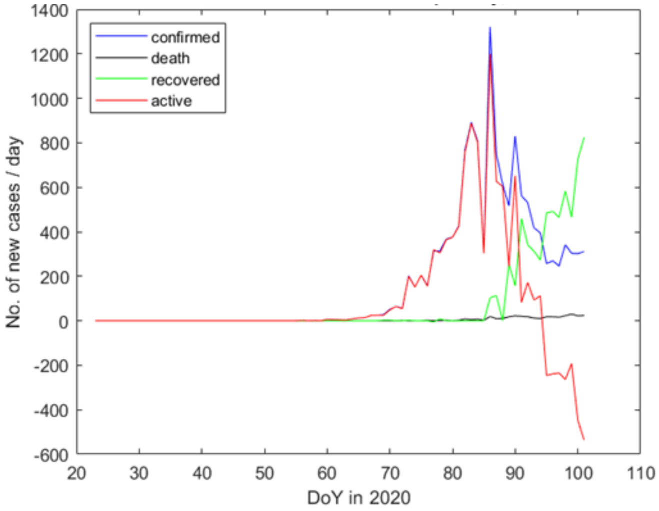

Figure 10: Time series of confirmed, fatal, recovered, and active cases on daily basis in Austria until 10 April 2020

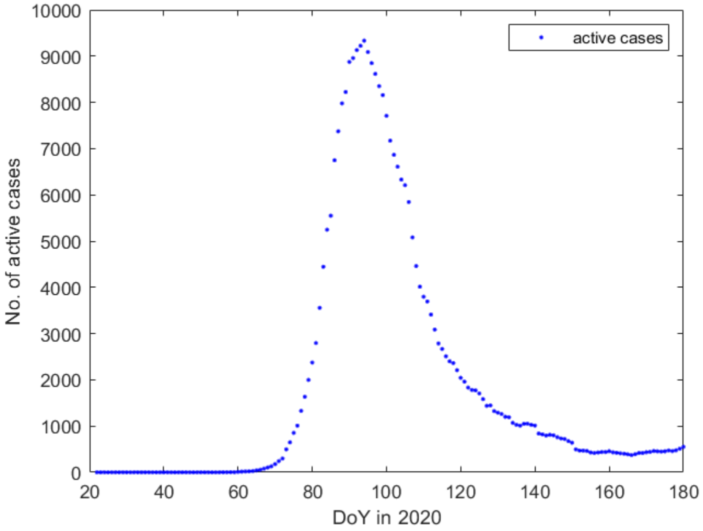

Figure 11: Active cases in Austria until 28 June 2020

When a country is just passing the maximum of active cases, not only the Gaussian curve but also the daily number of new confirmed cases should show a drop to probable change in tendency; otherwise the Gaussian bell curve may by unfounded. And this is tricky: there may be changes in the policies of reporting confirmed cases in certain countries as positive public sentiment can motivate the society to keep on following the restrictions, or just simply providing a political advantage of the ruling government to be effective in overcoming the virus within short time and few victims (compared to other countries). All in all, the data can be essentially distorted for several reasons, and the amount of distortion may change by time.

7. World-Wide Status by 28 June 2020

By 28 June 2020, the status is that the spread of COVID-19 is world-wide (global), and it is still in spreading phase. The situation is, however, different for the different continents. Asia shows a very diverse picture. Those countries, which were first affected (China, Hong Kong, Macau, Taiwan, Japan, South Korea) are definitely over the maximum of the 1st phase of the pandemic.

Actually, in these countries (apart from China) a smaller 2nd phase of the virus was starting by late May or early June, which turned to be much less virulent than the 1st phase was. Also, the Indochinese Peninsula (Myanmar, Thailand, Cambodia, Laos, Vietnam, Malaysia, Singapore) could over the critical phase of the virus. The rest of Asia, apart from some countries, which could (or seemingly could) successfully apply restrictions against it (Turkey, Georgia, Afghanistan, Tajikistan, Mongolia, United Arab Emirates, Qatar, Maldives, Sri Lanka, Bhutan, Brunei), the COVID-19 is still in spreading phase. Additionally, a relevant 2nd phase can be observed now in Israel, Kuwait, and Iran.

Europe has mostly exceeded the maximum of the first phase of the virus, apart from Sweden (which has declared certain resistance against social distancing) it is exceeding in France, Belgium, Moldova, Ukraine, Bulgaria, and Greece. By the end of June, the Balkan became a hot spot as the observed 2nd phase essentially exceeds the original one (Albania, Bosnia and Herzegovina, Montenegro, Kosovo, Macedonia). A general feature for Europe is the presence of an escalating 2nd phase (Spain, Switzerland, Czechia, Slovakia, Croatia, Romania, among others).

The majority of Africa is in extending phase, the virus is obviously spreading, apart from some countries (Western Sahara, Guinea, Burkina Faso, Niger, Chad, Uganda and Djibouti) and islands (countries of the Mascarenhas Archipelago, Mayotte). Note that in some cases decrease of the active cases was observed around a month ago, however a second phase of the spread is relevant now, which often exceeds the number of active cases of the first phase (Morocco, Seychelles, Tunisia). The virus in North America is undoubtedly in spreading phase in the USA, but in Canada and on some small islands (Saint Pierre and Miquelon, Bermuda) the maximum of the infection has been overcome. In South America, the situation is similar, apart from Uruguay and some island countries (Bonaire, Sint Eustatius and Saba, Aruba, Falkland Islands). In the continental Central America, it is also spreading, but most of the Lesser and Greater Antilles could get over the maximum (apart from the Dominican Republic, Puerto Rico and Martinique).

Australia has efficiently fought against the virus; on the continent and in New Zealand the 2nd phase has been observed to escalate starting by mid-June. In most islands of Oceania, it has not even been detected, but where the virus was observed, it has been overcome efficiently. Obviously, small islands (regardless their location along the globe) can get rid of the virus over much shorter period with much less observed cases. Due to their small size and population, in most cases no actual phases of the virus can be detected.

All in all, continent-wise analysis of the spread of the virus shows a diverse (or even a promising) picture. The reality is, however, not that nice. Among the 20 most populated countries, in India, USA, Indonesia, Pakistan, Brazil, Nigeria, Bangladesh, Mexico, Ethiopia, Philippines, Egypt, Congo and Iran it is obviously in extending phase. These countries give the 44% of the world's population. Additionally, among the most populated 20 countries the success in overcoming the virus in Russia is also not convincing, meanwhile China (without acknowledging a 2nd phase) has again started to isolate several megacities. So globally we face an unfavourable situation. Meanwhile, the spread of a 2nd phase of the virus seems to be unavoidable in the next period, already manifesting obvious traces of its emergence. The 2nd phase depends relevantly on the change and timing of prevention activity of a country. In some cases, it may just be a temporary increase and may not turn to be an actual 2nd phase of the disease, while other cases it may become the main phase. Everything depends on the international attitude to social distancing.

8. Conclusions and Summary

In the case of COVID-19, it has been already recognized that a very high percentage of infected people show no symptoms at all, thus without testing they cannot be recognized. An ugly feature of this virus is that they are still highly contagious. Testing policy is notably different worldwide, apart from Germany, no intensive testing in the whole society is done. (This has resulted in the high number of confirmed cases and very low death toll in Germany compared to other European countries). Probably, there are orders of magnitude differences in the actual and confirmed cases worldwide, so even the order of magnitude of actual cases is unknown. This makes all GIS analysis arguable. Therefore, all results of this study are arguable mainly due to the unreliability of the available data.

Still, the developed MATLAB program provides a tool for geostatistical analysis of the continuously developing time series of contagion, thus can be a useful tool for decisionmakers on planning future actions. As the situation is evolving, up-to-date analysis of the spread of the virus is going to be performed. The most important tools for monitoring and analysing the spread of the COVID-19 virus are the geographical display of the status of the virus spread, time series of confirmed / fatal / recovered / active cases of the disease on cumulative and on daily basis, centre and radius of the spread, best fitting bell curve of the active cases, and general statistics (CRF, modified CFR, attack rate, mortality rate). Each feature can be used globally or country-wise.

Results on 28 June 2020 indicates that the COVID-19 globally is in ascending phase yet, however, Europe, Australia and Oceania is on the right track to overcome the virus, while Africa, North America, Central America and South America are in clearly in the extending period, and Asia is diverse. Obviously, small islands (regardless their location along the globe) are rather dynamic; they can get rid of the virus over much shorter period with much less infections, but can quickly enter to and overcome on a 2nd phase of the virus.

Note, that all the information used for deriving these conclusions is based on reported data, which obviously shows essential distortions. For example, in case of North Korea, no confirmed cases were officially reported, which is nonsense. Or, in the cases of Cambodia, Laos and Vietnam, no fatal cases were recorded, which also seems to be unreliable with respect to the number of confirmed cases. Country by country analysis of the time series indicates that the reporting is not consistent at all, all three components of active cases show high variability. The number of confirmed cases highly dependent on the number of tests made. Also, the number of fatal cases is arguable as in some countries all deaths, where the person has been infected are considered to be due to COVID-19, while in other cases only the direct causes are considered. Particularly the reporting of recovered cases shows high inconsistency: in some countries it is presented on daily basis, in others it is very hectic, e.g. in case of Norway, after months of no recovers, suddenly 7695 recovered cases have been reported. Recovered cases also contains negative data, which is uninterpretable. These are all proofs that local political intentions and the quality health care system may harshly distort the analyses.

References

Buss, S. R. and Fillmore, J. P., 2001, Spherical Averages and Applications to Spherical Splines and Interpolation. *ACM Transactions on Graphics*, Vol. 20(2), 95-126. doi: https://doi.org/10.1145/502122.502124.

Day, M, 2020, Covid-19: Identifying and Isolating Asymptomatic People Helped Eliminate Virus in Italian Village, BMJ, 368:m1165. doi: 10.1136/bmj.m1165.

Dong, E., Du, H. and Gardner, L., 2020, An Interactive Web-Based Dashboard to Track COVID-19 in Real Time, Lancet Infect Dis, Published Online. doi: https://doi.org/10.1016/S1473-3099(20)30120-1.

Ioannidis, J. P. A., 2020, A Fiasco In The Making? As the Coronavirus Pandemic Takes Hold, We Are Making Decisions Without Reliable Data, STAT News. Link: https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/ [accessed 23 August, 2020].

Moré, J. J. and Sorensen, D. C., 1983, Computing a Trust Region Step. *SIAM Journal on Scientific and Statistical Computing*, Vol. 3, 553-572.

Nelder, J. A. and Mead, R., 1965, A Simplex Method for Function Minimization. *Computer Journal*, Vol. 7(4), 308-313. doi:10.1093/comjnl/7.4.308.

Panozzo, D., Baran, I., Diamanti, O. and Sorkine-Hornung, O., 2013, Weighted Averages on Surfaces. *ACM Transactions on Graphics*, Vol. 32(4), article No. 60. doi: https://doi.org/10.1145/2461912.2461935.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G. T., Fu, H., Dighe, A., Griffin, J. T., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunubá, Z., Fitz, J.

R., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Laydon, D., Nedjati-Gilani, G., Riley, S., van Elsland, S., Volz, E., Wang, H., Wang, Y., Xi, X., Donnelly, C. A., Ghani, A.C. and Ferguson, N. M., 2020, Estimates of the Severity of Coronavirus Disease 2019: A Model-Based Analysis. The Lancet Infectious Diseases, Vol. 20(6), 669-677. doi: https://doi.org/10.1016/S1473-3099(20)30243-7.

Wilder-Smith, A., Chiew, C. J. and Lee, V. J., 2020, Can We Contain the COVID-19 Outbreak with The Same Measures as for SARS?. *The Lancet Infectious Diseases*, Vol. 20(5), 102-107, doi: https://doi.org/10.1016/S1473-3099(20)30129-8.

Worldometer, 2020, Countries in the World By Population (2020). Online database, available at https://www.worldometers.info/world-population/population-by-country/. [accessed 23 August, 2020].