Adam M. Ćmiel[1], Bogdan Ćmiel[2]

[1] Institute of Nature Conservation, Polish Academy of Sciences, al. A. Mickiewicza 33, 31-120 Kraków, Poland

[2] Faculty of Applied Mathematics, AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Kraków, Poland

corresponding author: cmiel@iop.krakow.pl

A new, simple method of describing COVID-19 trajectory and dynamics in any country based on Johnson Cumulative Distribution Function fitting.

Abstract

This paper present simple method to study and to compare the infection dynamics between countries based on curve fitting to the publicly shared data of COVID-19 confirmed infections reported by them. Presented method was tested using data from 80 countries from 6 regions. We found that Johnson Cumulative Distribution Functions (CDF) are extremely well fitted to the data (R^2>0.99) and that Johnson CDF is much better fitted to the data at its tails than both commonly used Normal and Lognormal CDF. Fitted Johnson CDFs can be used to obtain basic parameters of the infection wave, such as the percentage of the population infected during the infection wave, day of the start, peak and the end of the infection wave, as well as the duration of the infections wave and the duration of the wave increase and decrease. These parameters may be easily biologically interpreted and used both in describing the infection wave dynamics and in further statistical analysis. The usefulness of the obtained parameters was demonstrated on two examples: the analysis of the relation of the Gross Domestic Product (GDP) per capita and the analysis of the population density on the percentage of the population infected during infection wave, the day of the start, and the duration of the infection wave in analyzed countries. We found that all of the abovementioned parameters were significantly dependent on the GDP per capita, while only the percentage of population infected was significantly dependent on the population density in analyzed countries. Also, if used with caution, presented method has some limited ability to predict the future trajectory and parameters of the ongoing infection wave.

Introduction

COVID-19 is a highly contagious disease, caused by the SARS-CoV-2 coronavirus. The virus was first detected in Wuhan (Central China) in December 2019, but as early as mid January, the virus quickly spread throughout China. On 13 January 2020, the first case outside China was confirmed and on 24 January, the first case in Europe was reported. In the second half of February 2020, outbreaks with hundreds of cases erupted in South Korea, Italy and Iran (Skórka et al., 2020) and COVID-19

was declared as a pandemic by the World Health Organization on March 11, 2020 (Ducharme, 2020).To date, globally, over 64 million infections and almost 1.5 million death cases were reported (WHO, 2020).

Since the very beginning of the pandemic, many models have been proposed to understand the outbreak dynamics of COVID-19 (e.g. IHME, 2020; UGSDSC, 2020; LANL, 2020; Ferguson et al., 2020; Kissler et al., 2020; Aleta et al.; Hellewell et al., 2020) and were used by policymakers (e.g. US Government) to allocate resources or plan interventions. Some of them, such as early IHME model received fair amount of criticism (Jewell et al., 2020). COVID-19 modelling studies generally follow one of two general approaches: forecasting models and mechanistic models; although there are hybrid approaches (Holmdahl and Buckee, 2020). Forecasting models are often statistical in nature, fitting a line or curve to data and extrapolating from there, without incorporating the process that produces the pattern (Holmdahl and Buckee, 2020), while mechanistic models simulate the outbreak through interacting disease mechanisms by using local nonlinear population dynamics and global mixing of populations (Hethcote, 2000). Purely statistic models are reliable only within a short time window and may be useful to make rapid short-term recommendations, whereas mechanistic modelling can be useful to explore how the pandemic would change under various assumptions and political interventions (Kuhl, 2020).

Since its beginning, COVID-19 pandemic generated huge amount of data and probably is the best documented disease in history. New cases, active cases, death cases, number of tests performed data are usually daily published by official sources (e.g. governments), gathered and publicly shared as freely accessible datasets (e.g. Hasell, et al. 2020). This makes a possibility for researchers to focus on analyzing the pandemic and its dynamics also in other fields than epidemiology. However, abovementioned models provide many pandemic parameters, useful in predicting different scenarios of future infections, day, probability and duration of future pandemic peaks, which is extremely useful for policymakers in planning interventions, however they may not be very useful in other fields than epidemiology. Thus the urgent need of developing methods of describing the trajectory of pandemic waves arose. Such methods should be easy to apply, and should provide parameters describing trajectory and dynamics of the epidemic, which are easy to interpret and to use in further statistical analysis by researchers from other fields (e.g. sociology, biology, ecology, etc.) which can deepen our understanding of the COVID-19 pandemic.

The aim of this paper is to present a new simple method based on curve fitting to the reported data on confirmed cases of infection, to study and compare the infection dynamics between countries (or regions). The method is based on the Johnson Cumulative Distribution Function (CDF) fitting, and was tested using data from 80 countries from 6 regions (Africa, Asia, Europe, Oceania and both North and South America). Also, Johnson CDFs were used to calculate basic parameters of the infection wave dynamics, such as the percentage of the population infected during the infection wave, day of the start, peak and end of the infection wave, as well as the duration of the infections wave and the

duration of the wave increase and decrease. This parameters are easy to interpret and may be used in further statistical analysis of epidemic dynamics, which was demonstrated on the examples of the influence of Global Domestic Product (GDP) per capita and the influence of population density on the percentage of infections and the day of the start and the duration of the first infection wave in analyzed countries. Both the presented method and techniques employed are all straightforward and well known and the purpose of the paper is to illustrate how simple techniques can be used to solve otherwise difficult problems, such as description of the epidemic wave.

Materials and methods

The data used in this study was obtained from Our World in Data COVID-19 dataset (Hasell, et al. 2020) from December 2019 to 19 November 2020. Presented method was tested on 80 countries from 6 Regions: 1) Africa (Democratic Republic of Congo, Egypt, Ethiopia, Kenya, Morocco, Nigeria, Somalia, South Africa, South Sudan, Sudan and Zimbabue), 2) Asia(Afghanistan, Bangladesh, Cambodia, China, India, Indonesia, Iran, Iraq, Israel, Japan, Lebanon, Myanmar, Pakistan, Philippines, Saudi Arabia, Singapore, South Korea, Sri Lanka, Syria, Taiwan, Thailand, Turkey, Vietnam), 3) Europe (Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czechia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom), 4) North America (Canada, Jamaica, United States of Mexico, United States of America) 5) Oceania (Australia, Fiji, New Zealand, Papua New Guinea), and 6) South America (Argentina, Bolivia, Brasil, Chile, Colombia, Paraguay, Peru, Uruguay, Venezuela).

In order to make the data comparable between countries, for each country, number of infections in each day of the pandemic, was standardized, and were presented as a percentage of the population of a given country infected (number of confirmed infections in a given country/country population*100%). Also, a five-days moving average was calculated using percentage of infections to smooth the data and to minimize the effect of lower number of tests performed and lower number of confirmed infections during some short periods (e.g. weekends).This makes the loss function more regular i.e. it has less relative extrema, which makes it easier to find global extremum. Nevertheless, all presented $R^2$for obtained Johnson CDFs are calculated using raw (not smoothed) data.

Fitting Johnson CDF by moments

Johnson (1949) described a system of frequency curves that represents transformations of the standard normal curve (detailed description in Hahn and Shapiro, 1967). Applying these transformations to a standard normal variable allows a unique distribution to be derived for whatever combination of mean, standard deviation, skewness, and kurtosis occurs for a given set of observed data. The standard method of fitting Johnson curves is to use four coefficients defining a Johnson distribution: two shape (γ, δ), a location (ξ), and a scale (λ) coefficient:

$$F(x) = \Phi(\gamma + \delta sinh^{-1}\left(\frac{x-\xi}{\lambda}\right)) \quad (1)$$

,where Φ is cumulative distribution function of standard normal distribution. However, this method is not intuitive (i.e. it is difficult to set starting points from the data to perform numerical fitting). Thus alternative method for fitting Johnson curves, using first four moments (mean, variance, skewness and kurtosis) of an empirical distribution was selected (detailed description in Hahn and Shapiro,1967 and Hill et al.,1976).All statistical fits in the paper were performed using the Levenberg-Marquardt algorithm (Moré, 1978) to solve the corresponding non-linear least square optimization problem. Convergence criterion was set to $1.0E^{-10}$.

Fitting Johnson CDF to the epidemic waves

There is no strict definition for what is or is not an epidemic wave or phase. The intuitive definition of the pandemic wave traces the development of an epidemic over time and/or space. During an epidemic the number of new infected cases increases (often rapidly) to a peak and then falls (usually more gradually) until the epidemic wave is over.

The epidemic dynamics may highly differ between countries. Since the beginning of the pandemic, in some countries only one epidemic wave was observed (e.g. Afghanistan, Argentina), in some countries two epidemic waves were observed (e.g. Australia), while in others even more epidemic waves were observed, which also may overlap and interfere each other (e.g. Croatia, where four overlapping and interfering waves were observed). Also, in many countries, a range of various levels of the lockdown were applied to slow down or "flatten" the infection curve, the epidemic waves may not follow the Farr's law (which states that epidemics tend to rise and fall in a roughly symmetrical pattern or bell-shaped curve) and may be asymmetrical.

The basic assumption is that each epidemic wave *W* in a given country may be described by a five parameters scaled Johnson CDF: scale parameter (*s*), and abovementioned moments: expected value (mean; *E*), variance (*V*), skewness (*S*) and kurtosis (*K*)

$$W(t)=s*F_{E,V,S,K}(t) \quad (1)$$

, where *t* is the time measured since the day of the beginning of the pandemic and function $F_{E,V,S,K}$ is Johnson CDF with parameters γ, δ, ξ, λ assuring mean, variance, skewness and kurtosis equal to *E,V,S,K* respectively (see Hahn and Shapiro,1967; Hill et al.,1976). The *S* and *K* parameters were expected to improve the curve fit at the tails of the epidemic wave in case it was not symmetrical or heavy tailed.

Obtaining basic epidemic wave parameters and their biological interpretation

Once the Johnson CDFs were fitted to each pandemic wave in a given country, basic parameters obtaining the wave dynamics: (1) 2.5% quantile ($Q_{2.5\%}$), (2) 50% quantile (median; $Q_{50\%}$), (3) 97.5% quantile ($Q_{97.5\%}$) were calculated:

$$Q_{2.5\%} = F^{-1}_{E,V,S,K}(2.5\%) \quad (2)$$

$$Q_{50\%} = F^{-1}_{E,V,S,K}(50\%) \quad (3)$$

$$Q_{97.5\%} = F^{-1}_{E,V,S,K}(97.5\%) \quad (4)$$

The disadvantage of fitting Johnson curve by its moments is that it is not possible to determinate its mode analytically. Thus the mode of each Johnson CDF was determined numerically:

$$M = \arg\max f_{E,V,S,K}(x) \quad (5)$$

, where $f_{E,V,S,K}$ is Johnson Probability Density Function (PDF).

The obtained parameters have an intuitive biological interpretation (Fig. 1): the scale parameter (*s*) indicate the total percentage of infections during a given epidemic wave (P_{inf}), $Q_{2.5\%}$ indicate the day when infection wave starts, while $Q_{97.5\%}$indicate its end. Median ($Q_{50\%}$) indicate the day when the half of the total percentage of infected during a given wave was reached. Finally, the mode (*M*) indicate the day of the peak occurrence. Additionally, one can easily obtain the wave duration (*T*)

$$T = Q_{97.5\%} - Q_{2.5\%} \quad (5)$$

the duration of wave increase (t_i)

$$t_i = M - Q_{2.5\%} \quad (6)$$

and the duration of the wave decrease (t_d)

$$t_d = Q_{97.5\%} - M \quad (7)$$

Also, the parameter measuring the asymmetry of the infection wave (*A*) can be easily obtained as a ratio

$$A = t_i/t_d \quad (8)$$

All of the abovementioned parameters may be easily used in further statistical analysis, which was shown on examples: 1) the relationship between Gross Domestic Product (GDP)per capita and basic parameters describing the dynamics of the first wave of infections: *M*, *T*, and P_{inf}, and 2) the relation between population density and basic parameters describing the dynamics of the first wave of infections: *M*, *T*, and P_{inf}. Only first wave of infections in each country was taken into account, because in some countries, second (and consecutive) waves were not observed, and they would have been excluded from the analysis.

Comparing curves: Johnson vs Normal and Lognormal CDF

The differences between Johnson, Normal and Lognormal CDF were presented on the data from Afghanistan, where only one epidemic wave was observed. The differences were shown by comparing the R^2,P_{inf}, $Q_{2.5\%}$, *M*, and $Q_{97.5\%}$ parameters. Both 2.5% and 97.5% quantiles for normal and lognormal distributions, were obtained using inverse Normal and inverse Lognormal PDF respectively.

Fitting Johnson CDF to the ongoing wave and possibility of prognosis

Fitting Johnsons curve to the ongoing wave result in obtaining parameters, which can also be interpreted as a prognosis of the future shape and dynamics of infection wave. In such case, P_{inf}, M and $Q_{97.5\%}$ indicate predicted percentage of infections, predicted day of the peak and predicted day of the end of the ongoing wave respectively, which also can be used to calculate predicted time of increase, decrease and duration of the ongoing infection wave. Because presented method is intended to describe infection dynamics rather than predicting its future outcome, the accuracy of the prognosis was presented only on the data on the first wave of infection observed in the United Kingdom in the Supplementary Materials.

Examples of application

The relation between Gross Domestic Product (GDP) per capita and the relation between population density and the dynamics of the first wave of COVID-19 infections

The data on the GDP per capita and population density in 80 analyzed countries were obtained from Our World in Data COVID-19 dataset (Hasell, et al. 2020).

The relationship between GDP per capita and the relation between population density and basic parameters describing the dynamics of the first wave of infections (M, T, and P_{inf}) obtained using presented method of Johnson CDF fitting was tested using the quantile dependence function method, which was described in detail in Ćmiel and Ledwina (2020).This method was designed for measuring, visualizing the dependence structure, and testing of independence of two random variables. It exploits a recently introduced local dependence measure (quantile dependence function q), which gives a detailed picture of the underlying dependence structure and provides a means to carefully examine the local association structure at different quantile levels (Ćmiel and Ledwina 2020).

Results

The examples of fitted Johnson curves to the data from countries where one ongoing infection wave (Argentina), one infection wave (Afghanistan), two infection waves (Australia) and four overlapping and interfering infection waves (Croatia) were observed was presented at Fig. 2.Fitted four Johnson CDFs to the four waves of infections observed in Croatia, with areas where waves are overlapping and interfering was presented in detail at Fig. 3A.

Johnson CDF fitting tested using data obtained from 80 different countries showed that all curves were extremely well fitted: the lowest R^2 obtained was 0.995 (Fiji), while the highest R^2 was 0.99997 (Iraq), while the mean and median R^2 was 0.9995 and 0.9997 respectively. Fitted functions with R^2 and COVID-19 trajectory plots with fitted functions for each country were presented the Supplementary Materials (Table S1; Figure S1-S6).

Fitting Johnson, Normal and Lognormal distribution curves to the single wave of infection observed in Afghanistan showed, that the best fitted was the Johnson CDF (R^2=0.9998), while both Normal (R^2=0.9980) and Lognormal (R^2=0.9989) distributions were worse fitted, mainly at the tails of

the infection wave (Fig. 3B). Obtained parameters $Q_{2.5\%}$, M, $Q_{97.5\%}$ for the infections wave in Afghanistan using Johnson CDF fitting were 59, 100, 209 respectively, while the same parameters obtained using Normal CDF fitting and Lognormal CDF fitting were 57, 105, 152 and 65, 98,167 respectively.Percent of confirmed population infected during the infection wave obtained using scale parameters (s) of fitted Johnson, Normal and Lognormal distributions were 0.1028%, 0.0984% and 0.0997% respectively.

Among analyzed countries, 17 (21.3%) countries were described by fitting one wave of infections, 35 (43.8%) countries were described by fitting two waves of infections, 24 (30%) countries were described by fitting three waves of infections and 4 (5%) countries were described by fitting four waves of infections (Table S1).

The basic statistics for the obtained skewness parameters of Johnson distributions fitted to the first pandemic waves in 80 analyzed counties showed, that in majority of them, the first wave of infection was skewed (median S=1.5; minimum S=0; maximum S=141.5). First wave of infection was symmetrical in 16 countries (20%;A<1.05). Also, basic statistics for parameter A showed, that time of wave decreasing is longer than time of wave increase (mean A= 4.7; median A=2.9; minimum A=1.0; maximum A=22.4).

The results of the analysis of the associations between GDP per capita and M, T and P_{inf} parameters showed, that the percentage of confirmed infections during the first epidemic wave in analyzed countries was dependent on the GDP per capita (p=0.0147; Fig 4A), as well as the time of the peak occurrence (M; p=0.0002; Fig. 4B) and the duration of the first epidemic wave (T; p=0.0087; Fig. 4C).The relation between the percentage of infections and GDP per capita showed rather global positive dependence (Fig. 4A), which means that the higher GDP per capita, the higher percentage of infections during the first epidemic wave. The relation between the time of peak occurrence and GDP per capita showed local negative dependence for countries where peak occurs late (above median; Fig. 4B) which means that the very early occurrence of peak is rather not correlated with GDP per capita but in case when the peak does not occur early the higher GDP per capita, the earlier peak occurs. The similar relation was also observed for the relation between the duration of the infection wave and GDP per capita (Fig. 4C), i.e. the very short duration of the first epidemic wave is rather not correlated with GDP per capita but in case when the duration of the first epidemic wave is not short, the higher GDP per capita, the shorter first epidemic wave.

The results of the analysis of the associations between population density and M, T and P_{inf} parameters showed that the percentage of infections during the first epidemic wave in analyzed countries was dependent on the population density (p=0.0079; Fig 4D), while the day of the peak occurrence and the duration of the first epidemic wave were not dependent on population density (T: p=0.4243; Fig. 4E; M: p=0.5924; Fig. 4F).The relation between percentage of infections and population density showed local negative dependence (Fig. 4D) e.g. in case when population density is

not very high but the percentage of infections is rather high. In such case the higher population density the lower percentage of infections.

Discussion

The method presented in this paper gives an indication of the spread of the COVID-19 disease in particularly any country, which provides daily numbers of infected cases. Both the presented method and techniques employed are all straightforward, well known and easy to use, since Johnson CDF fitting is available in many statistical/calculus packages, e.g. R, Statistica, MATLAB, MS Excel. Using alternative method of fitting using moments instead of shape, location and scale parameters makes it easier to set starting points for numerical fitting (e.g. by visual analyzing the scatter plot of number of infected in time). Obtained curves are extremely well fitted, which was shown on the example of 80 different countries from 6 regions. Also, obtained parameters are easy to interpret and ready to use in further analysis, such as finding associations between them and other variables which may be associated with COVID-19 dynamics, i.e. GDP per capita, population density.

To date, some research used curve-fitting with a Normal distribution to answer the real time request and applied it to COVID-19 in Wuhan (Tomie 2020) since it was known that flu epidemic followed a Normal distribution, whereas other researchers noticed the COVID-19 profile has a feature to leave a trail in an asymmetric and applied a Lognormal distribution curve fitting (Nishimoto and Inoue 2020). The results presented in this paper showed, that in 79% of analyzed countries, first wave of infections were highly skewed, which suggest that unlike the flu, COVID-19 epidemic does not follow the normal distribution and should not be modelled in this manner. In such case log-normal distribution fitting seems to be better, however, as it was presented on the example of Afghanistan, the differences in R^2 between Johnson, Normal and Lognormal CDFs seem to be small, but the difference is ca. 1 level of magnitude in favour of Johnson CDF. Moreover, one can see, that both Normal and Lognormal CDFs are fitted worse at the tails of the infection wave than Johnson CDF (Fig. 3B), and both showed lower number of infections than it was observed (raw data) and lower than obtained using Johnson CDF. Also, fitted Lognormal curve starts to increase later than Normal and Johnson distribution curves, which in consequence would led to incorrect estimation of the beginning of the wave (11 days later than it was obtained using Johnson distribution), whereas Normal distribution is far worse fitted at the right tail than Johnson and Lognormal distributions, because the wave of infections observed in Afghanistan was not symmetrical. Beside that using Normal distribution would unable estimating the true duration of the wave decrease (it is equal to the time of the wave increase by definition), it also leads to the much lower estimation of the day when the wave of infections ends (57 days earlier than estimated using Johnson distribution), which is caused by "too fast" flattering of the Normal CDF (Fig. 3B). Extremely high R^2 obtained for 80 analyzed countries (Supplementary Materials) suggest that Johnson curves class is flexible enough to almost perfectly follow the course of the epidemic in this countries. This results from the fact, that both skewness and kurtosis are estimated

parameters during Johnson curve fitting procedure, whereas the shape of other commonly used curves (Normal, Lognormal, Weibull) is more or less imposed. This result also suggests that Johnson distribution should be preferred in curve-fitting approach for COVID-19 data.

Presented curve fitting method was designed primarily to obtain easy in interpretation parameters describing past trajectory of COVID-19 infection, but parameters describing actually ongoing wave of infection, especially in its early stage (before the peak), may be interpreted as a forecast of future course of the pandemic. However, in such case, extreme caution is advised (see Jewell et al. 2020). Presented method is purely statistical model and it does not incorporate the process that produces the number of infections pattern, and does not account for any parameters governing transmission, disease, and immunity. Also, curve fitting techniques cannot predict the occurrence of future peaks. Thus, for long term prognosis and modelling the future scenarios of the pandemic, it is recommended to use more reliable methods, based on SEIR models. Nevertheless, some short term prognosis can be obtained using presented method, which may be useful for policymakers in rapid, short term intervention planning, however one must keep in mind the abovementioned limitations of presented method, as well as the limitations resulting from the data colleting and reporting, which are discussed later in this section.

The results obtained in the presented example of the application of parameters describing COVID-19 dynamics showed, that the higher the GDP per capita, the higher percentage of the population infected was observed. This is quite unexpected result, however consistent with the result which was very recently reported by Liu et al. (2020), who found the positive correlation between human development index (HDI) and risk of infections and deaths of COVID-19 in Italy. Other obtained results showed that, excluding countries where peak of infections occurred very early and its duration was short, the higher GDP per capita, the earlier peak occurs and the first epidemic wave is shorter. This result, in turn, is similar to another very recent paper, which reported that the date of first CoVID-19 cases co-varies positively with GDP across countries, most probably due to their more intensive participation of the global tourism and traffic industries (Jankowiak et al. 2020). The other example showed that the higher population density the lower the percentage of population infected during first wave of infections. This also seems to be unexpected, however, a negative dependence result from fact that the infections are presented as a percentage, which does not scale proportionally with the population density. Another possible explanation is that in countries with high population density (e.g. China, Singapore), very strict (full) lockdowns were immediately applied (China, Kretschmer and Yang, 2020; Singapore, Cheong 2020), which could result in lower percentages of infected population than in countries with lower population density, where partial lockdown or no lockdowns at all were applied. Moreover, some research report positive correlation between population density and number of infections and related mortality (e.g. in India; Bhadra et al., 2020), while other report no evidence that population density is linked with COVID-19 cases and deaths (e.g. in USA; Carozzi et al., 2020). Nevertheless, presented examples showed the usefulness of the

presented method, but also, the very recent papers of Liu et al. (2020) and Jankowiak et. al. (2020) showed that the field of research on COVID-19, other than purely epidemiological modelling of the future pandemic scenarios, is rising, which indicate that the simple methods of obtaining parameters describing the infection waves, such as presented in this paper, may be very useful and can help to deepen our understanding of the COVID-19 pandemic.

The last but not least issue which has to be addressed is a key limitation in understanding of the COVID-19 pandemic, that the true number of infections is not known and the only known infections are those confirmed by tests. Moreover, testing strategies differs between counties i.e. in some countries only symptomatic cases are tested, while in other mass testing is performed. Also, most COVID-19 cases are asymptomatic and remain unreported (Peirlinck et al. 2020). Because of that, mortality data are generally considered as more reliable than testing-dependent confirmed case counts and used in COVID-19 epidemic modelling (e.g. Chikobvu and Sigauke, 2020). However, some countries only report COVID-19 deaths occurring in hospitals, whereas other report COVID-19 deaths when test has confirmed the infection (this makes number of death data testing-dependent as well). On the other hand, when laboratory diagnosis is not required (e.g. United Kingdom, UK Guidance), it is possible that other diseases reassembling COVID-19 symptoms may be reported as COVID-19 cause of death. It may also be difficult to evaluate the cause of death in cases, when patient had other disease (e.g. advanced stage of cancer) together with COVID-19. Taking all of the above into account, it is very likely that real number of deaths is also higher than the reported number of deaths, which was noticed in some countries (e.g. Italy, Foresti 2020, Stancati and Sylvers, 2020; China, Long et al., 2020). It seems that both confirmed new cases and confirmed deaths may not be reliable, but on the other hand, no other data is available. Some models (e.g. IHME 2020) are able to estimate true number of infections, but it is related to a number of additional assumptions, and is partly based on the reported testing-dependent data. Also, the relation between true number of infections and number of death is not well studied to date and require a number of assumptions. Using the number of infections seems to be the easiest way of obtaining basic data on the COVID-19 infection dynamics in a given country, as long as one is aware that publicly shared data show number of confirmed cases instead of number of real infections and takes this into account when interpreting the results.

In conclusion, presented method based on Johnson CDF curve fitting to the cumulative number of confirmed cases is straight forward, well known and easy to use. It provides curves which are extremely well fitted to the data, and obtained basic parameters of COVID-19 infection dynamics are easy to interpret and to use in further statistical analysis by researchers from other fields than epidemiology (e.g. sociology, biology, ecology, etc.), and can deepen our understanding of the COVID-19 pandemic. It also may be useful in short term prognosis, however, in such case caution is advised.

Acknowledgements

Both authors equally contributed to the study and are listed alphabetically. This study was financed partly by the statutory funds of the Institute of Nature Conservation, Polish Academy of Sciences and partly by the statutory funds of the Faculty of Applied Mathematics, AGH University of Science and Technology. We thank Magdalena Lenda and Piotr Skórka for their useful comments and suggestions.

References

Bhadra, A., Mukherjee, A., Sarkar, K. 2020. Impact of population density on Covid-19 infected and mortality rate in India. Modeling Earth Systems and Environment. https://doi.org/10.1007/s40808-020-00984-7

Carozzi F., Provenzano S., Roth S. 2020. Urban density and COVID-19, discussion paper series, IZA Institure of Labor economics, IZA DP No. 13440.

Cheong D. 2020. Coronavirus: Most workplaces to close, schools will move to full home-based learning from next week, says PM Lee. The Straits Times. Singapore. 3 April 2020.Accessed at https://www.straitstimes.com/singapore/health/most-workplaces-to-close-schools-will-move-to-full-home-based-learning-from-next on 29 November 2020.

Chikobvu D., Sigauke C. 2020. Statistical distribution fitting to the number of COVID-19 deaths in South Africa. Preprint, DOI: 10.21203/rs.3.rs-32411/v1

Ćmiel B., Ledwina T. 2020. Validation of association. Insurance: Mathematics and Economics 91: 55-67. https://doi.org/10.1016/j.insmatheco.2019.12.003

Johnson, N. L. 1949. Systems of Frequency Curves Generated by Methods of translation. Biometrika36: 149–176.

Ducharme J. 2020. TheWHOJustDeclaredCoronavirusCOVID-19a Pandemic. Time 2020;11March2020. Accessed at https://time.com/5791661/who-coronavirus-pandemic-declaration/ on 28 November 2020.

Ferguson N.M., Laydon D., Nedjati-Gilani G., Imai N., Ainslie K., Baguelin M., Bhatia S., Boonyasiri A., Cucunubá Z., Cuomo-Dannenburg G., Dighe A., Dorigatti I., Fu H., Gaythorpe K., Green W., Hamlet A., Hinsley W., Okell L.C., van Elsland S., Thompson H., Verity R., Volz E., Wang H., Wang Y., Walker P.G.T, Walters C., Winskill P., Whittaker C., Donnelly C.A., Riley S., Ghani A.C. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College COVID-19 Response Team. Accessed at https://doi.org/10.25561/77482 on 29 November 2020.

Foresti C.C.L. 2020. The real death toll for Covid-19 is at least 4 times the official numbers. Corriere Della Serra. 26 March 2020. Accessed at www.corriere.it/politica/20_marzo_26/the-real-death-toll-for-covid-19-is-at-least-4-times-the-official-numbers-b5af0edc-6eeb-11ea-925b-a0c3cdbe1130.shtml?refresh_ce-cp. on 28 November 2020.

Hahn G.J., Shapiro S.S. 1967. Statistical Models in Engineering. Wiley series on systems engineering and analysis, Wiley, ISSN 0084-019X, pp. 199-220.

Hasell J., Mathieu E., Beltekian D., Macdonald B., Giattino C., Ortiz-Ospina E., Roser M., Ritchie H. 2020. A cross-country database of COVID-19 testing. Scientific Data 7, 345.https://doi.org/10.1038/s41597-020-00688-8

Hellewell J., Abbott S., Gimma A., Bosse N.I., Jarvis C.I., Russell T.W., Munday J.D., Kucharski A.J., Edmunds J., Sun F., Flasche S., Quilty B.J., Davies N., Liu Y., Clifford S., Klepac P., Jit M., Diamond C., Gibbs H., van Zandvoort K., Funk S., Eggo R.M. 2020. Feasibility of controllingCOVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health, 8:e488-e496. https://doi.org/10.1016/S2214-109X(20)30074-7.

Hethcote H.W. The mathematics of infectious diseases. SIAM Review, 42: 599-653

Hill I.D., Hill R., Holder R.L. 1976. Algorithm AS 99: Fitting Johnson Curves by Moments. Applied Statistics 25: 180-189. https://doi.org/10.2307/2346692

Holmdahl S.M., Buckee C. 2020. Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us. The New England Journal of Medicine, 383:303-305. DOI:10.1056/NEJMp2016822

IHME. 2020. Institute for Health Metrics and Evaluation COVID-19 Projections. Accessed at https://covid19.healthdata.org on 29 November 2020.

Jankowiak Ł., Rozsa L., Tryjanowski P., Møller A.P. 2020. Strong negative covariation between toxoplasmosis and CoVID-19 at a global scale: a spurious indirect effect? Scientific Reports 10: 12512 (2020). https://doi.org/10.1038/s41598-020-69351-x

Jewell N.P, Lewnard J.A., Jewell B.L. 2020. Caution Warranted: Using the Institute for Health Metrics and Evaluation Model for Predicting the Course of the COVID-19. Annals of Internal Medicine, 173: 226-227. https://doi.org/10.7326/M20-1565

Kissler S.M., Tedijanto C., Goldstein E., Grad Y.H., Lipsitch M. 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science 22 May 2020: Vol. 368, Issue 6493, pp. 860-868 DOI: 10.1126/science.abb5793

Kretschmer F., Yang W. 2020. Wuhan lockdown: China takes extreme measures to stop virus spread.. Deutsche Welle. 23 January 2020. Accessed at https://www.dw.com/en/wuhan-lockdown-china-takes-extreme-measures-to-stop-virus-spread/a-52120126 on 29 November 2020.

Kuhl E. 2020. Data-driven modeling of COVID-19—Lessons learned. Extreme Mechanics Letters40,100921. https://doi.org/10.1016/j.eml.2020.100921

LANL. 2020. Los Alamos National Laboratory COVID-19 Cases and Deaths Forecasts. Accessed at https://covid-19.bsvgateway.org. on 28 November 2020.

Liu K., He M., Zhuang Z., He D., Li H. 2020. Unexpected positive correlation between human development index and risk of infections and deaths of COVID-19 in Italy. One Health, 10: 100174. https://doi.org/10.1016/j.onehlt.2020.100174.

439 Long Q., Siu-fung L., Mudie L. Estimates Show Wuhan Death Toll Far Higher Than Official Figure.
440 Radio Free Asia. 27 March 2020. Accessed at www.rfa.org/english/news/china/wuhan-deaths-
441 03272020182846.html. on 28 November 2020.
442 Moré J.J. 1978. The Levenberg-Marquardt algorithm: Implementation and theory. In: Watson G.A.
443 (ed.) Numerical Analysis. Lecture Notes in Mathematics, 630. Springer, Berlin, Heidelberg.
444 Nishimoto Y., Inoue K. Curve-fitting approach for COVID-19 data and its physical background.
445 medRxiv 2020.07.02.20144899. doi: https://doi.org/10.1101/2020.07.02.20144899
446 Peirlinck M., Linka K., Costabal F.S., Bhattacharya j., Bendavid E., Ioannidis J.P.A., Kuhl E.
447 2020.Visualizing the invisible: The effect of asymptomatic transmission on the outbreak dynamics
448 of COVID-19. Computer Methods in Applied Mechanics and Engineering 372: 113410.
449 Skórka P., Grzywacz B., Moroń D., Lenda M. 2020. The macroecology of the COVID-19pandemicin
450 the Anthropocene. PLoS ONE 15:e0236856.https://doi.org/10.1371/journal.pone.0236856
451 Stancati M., Sylvers E. 2020. Italy's Coronavirus Death Toll Is Far Higher Than Reported. The Wall
452 Street Journal. 1 April 2020. Accessed at www.wsj.com/articles/italys-coronavirus-death-toll-is-
453 far-higher-than-reported-11585767179. on 28 November 2020.
454 Tomie T. Understanding the present status and forecasting of COVID—19 in Wuhan. medRxiv
455 2020.02.14. https://doi.org/10.1101/2020.02.13.20022251
456 UGSDSC. 2020. University of Geneva and Swiss Data Science Center COVID-19 Daily Epidemic
457 Forecasting. https://renkulab.shinyapps.io/COVID-19-Epidemic-Forecasting
458 UK Guidance. Guidance for doctors completing Medical Certificates of Cause of Death in England
459 and Wales. HM Passport Office, Office fir National Statistics. Accessed at
460 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/8
461 77302/guidance-for-doctors-completing-medical-certificates-of-cause-of-death-covid-19.pdf. on 28
462 November 2020.
463 WHO 2020. World Health Organization Coronavirus Disease (COVID-19) Dashboard. Accessed at
464 https://covid19.who.int/ on 4 December 2020.
465

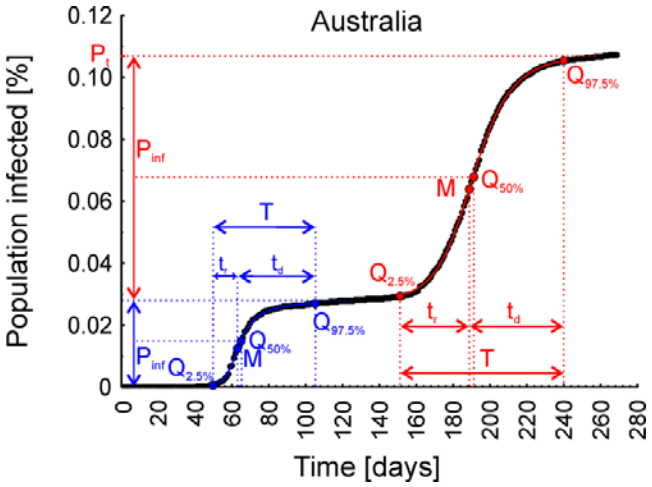

Figure 1. Graphical presentation of the interpretation of the obtained parameters from Johnson Cumulative Distribution Function fitting, describing the dynamics of the two infection waves observed in Australia. P_{inf} indicate the total percentage of infections in a given infection wave, $Q_{2.5\%}$ indicate the day when the infection wave starts, $Q_{97.5\%}$indicate the day when the infection wave ends $Q_{50\%}$ indicate the day when the half of the total percentage of infected during a given wave was reached, M indicate the day of the peak occurrence, T indicate the wave duration, t_i indicate the duration of the wave increase, t_d indicate the duration of the wave decrease, P_t indicate the total percentage of population infected after two waves of infections.

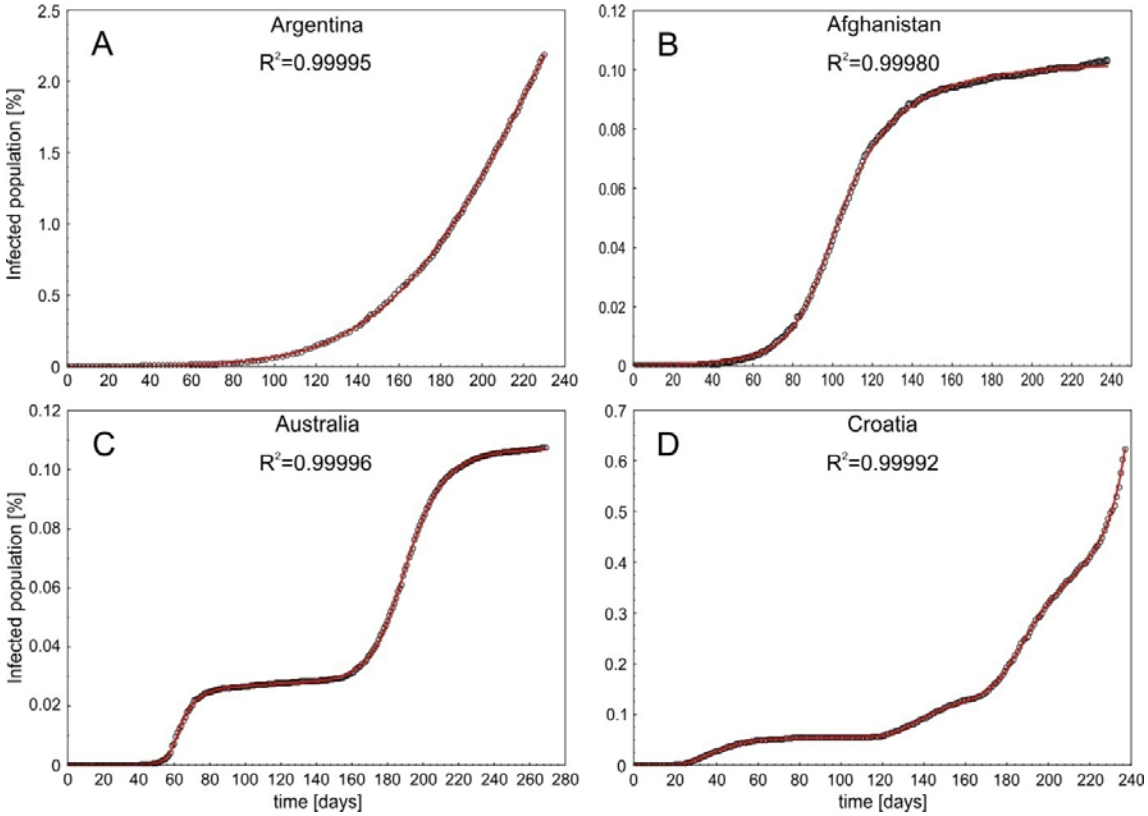

Fig. 2. Examples of fitted distributions in four scenarios of COVID-19 infection dynamics. A - one ongoing infection wave (before the peak), B - full one wave, C - two waves and D - four overlapping and interfering waves. Open dots indicate raw data, red lined indicate fitted Johnson Cumulative Distribution Functions.

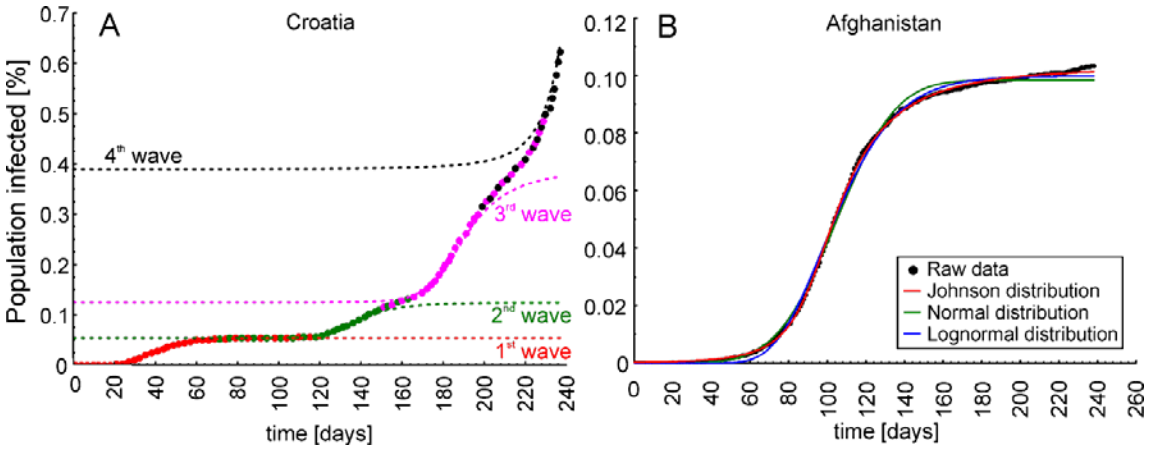

Fig. 3.A - the trajectory of four Johnson Cumulative Distribution Functions fitted to the four waves of infections observed in Croatia, with areas where waves are overlapping and interfering. B - the differences between fitted Johnson (red line) Normal (green line) and Lognormal Cumulative Distribution Functions to the raw data from Afghanistan (black dots).

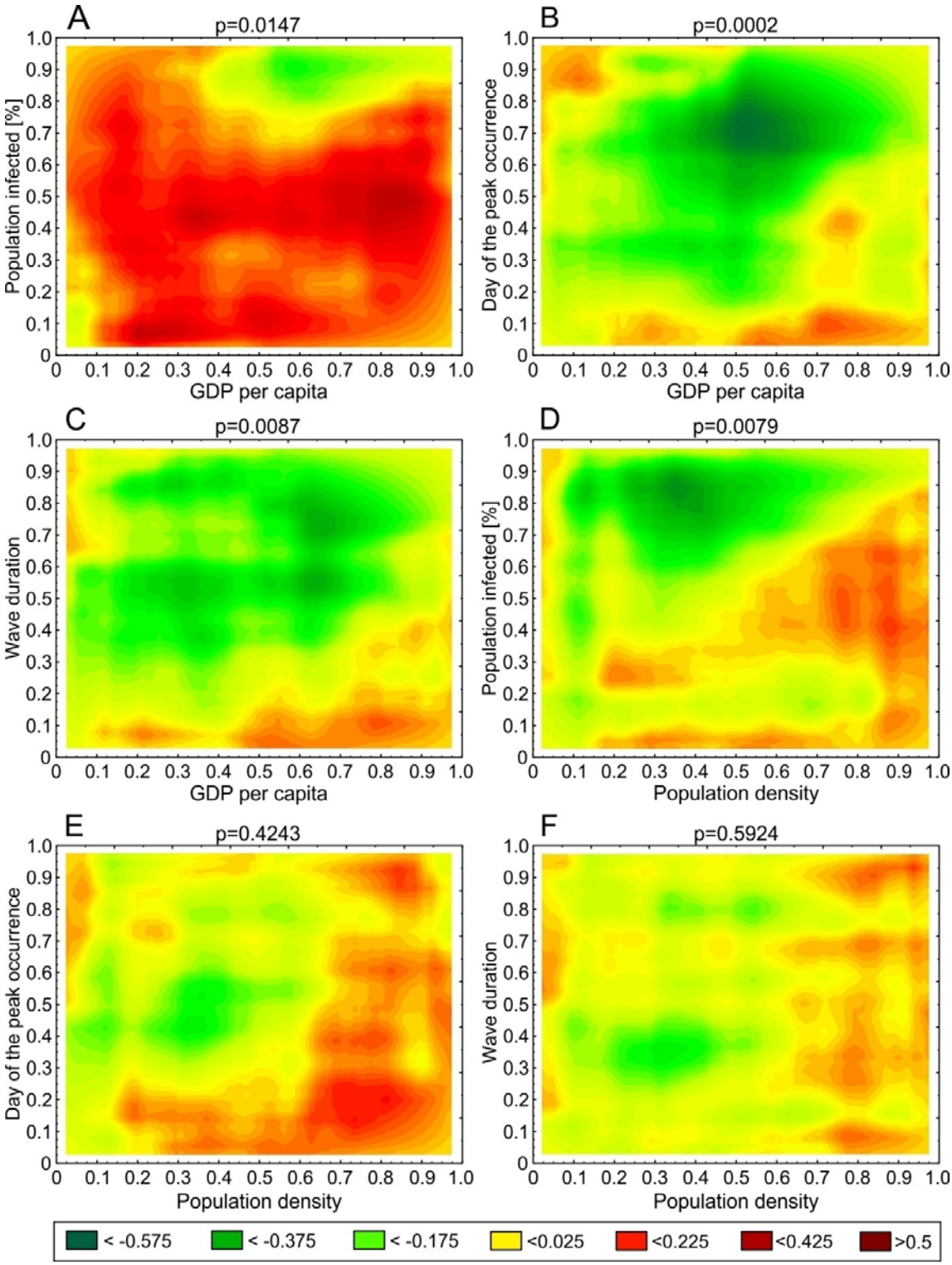

Fig. 4. Heat maps showing the local association structure between variables at different quantile levels obtained using quantile dependence function *q*.