Open Access

Advances in Environmental and Engineering Research

LIDSEN Publishing Inc.

Research Article

Modelling the Cumulative Number of COVID-19 Cases

Mieczysław Szyszkowicz *

112 Four Seasons, Ottawa, Canada; E-Mail: mszyszkowicz@yahoo.ca

* **Correspondence:** Mieczysław Szyszkowicz; E-Mail: mszyszkowicz@yahoo.ca

Academic Editor: Marco Bortolini and Francesco Gabriele Galizia

Special Issue: Energy Efficiency in Flexible and Reconfigurable Manufacturing: Emerging Trends, Models and Applications in the Industry 4.0 Era

Adv Environ Eng Res
2021, volume 2, issue 2
doi:10.21926/aeer.2102014

Received: April 14, 2021
Accepted: June 08, 2021
Published: June 10, 2021

Abstract

Each country has its own characteristics of COVID-19 infection trajectory and epidemic waves. Differences in government-implemented restrictions and social regulations result in variability of the virus transmissions and spread dynamics. This in turn results in various shapes of the growth function used to represent and describe the propagation of infection. Statistical methods are applied to fit non-linear functions to represent daily time-series data of the cumulative numbers of COVID-19 cases. The aim of this work is to fit various statistical models to the cumulative number of COVID-19 cases. Also to overview various types of the existed numerical methodologies. The data (since December 31, 2019) are available for almost each country in the world. As the examples, we used daily time-series data of the cumulative number of COVID-19 cases in Poland, Italy, Canada, and the USA. Non-linear approximations are applied to represent these time series data. The fitted functions allow us to investigate the dynamics of the pandemic. The constructed approximations are compositions of a few nonlinear functions, which describe the growth process of the COVID-19 infection trajectories. Two Gompertz functions and cumulative distribution functions (cdf) were estimated for the data of Poland and Italy (using the cdf for the normal distribution) and for the data of Canada and the USA (using the cdf for the gamma distribution). An analytical (parametric) functions

representation of the number of COVID-19 cumulative cases is a useful tool to study the propagation of epidemics.

Keywords
Cumulative cases; corona virus; distribution; epidemics; function; growth; infection

1. Introduction

This work is a contribution to the analysis of the spread of the coronavirus (SARS-CoV-2 or COVID-19) in different countries. The new infectious disease of coronavirus (COVID-19) in recent times is recognized as the most urgent and attractive field of research. To estimate the trend of infections in the world is very important task. Also to understand and describe the dynamics of this disease is important, for the current time and historical description. This study is needed and the primary benefits provided by the study is to describe (as a parametric function) and visualize the dynamics of the cases propagation. The World Health Organization (WHO) declared COVID-19 a pandemic on March 11, 2020 and many countries put in place various regulations and strict lockdowns. The declared pandemic and the regulations following this decision have huge impacts and consequences not only in the area of health, but also to other aspects of life, such as the economic, education, cultural and sporting activities, tourism and travel, and others. These impacts and effects with various intensities are observed around the world. The COVID-19 pandemic has altered routine life across the world, infecting hundreds of millions and killing over a million people as of September 2020 [1].

From a biological point of view, the virus that causes the coronavirus infection (COVID-19) is in a family of viruses *coronaviridae.* It results in a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It is a new virus, and has not yet been adequately studied to provide all the information on its characteristics and spread behaviour. The main clinical symptoms of the disease caused by SARS-CoV-2 virus, are fever, fatigue, dry cough, and difficulty breathing. Other symptoms might also be present, such as loss of smell and taste, conjunctivitis, and discoloration of fingers or toes (see for example https://www.healthline.com/health/coronavirus-symptoms#symptoms).

Among the important aims in the research on this disease is to model and understand the spread trajectories of COVID-19. The constructed mathematical models should well describe the pandemic behaviours, reconstruct the dynamics, provide a good approximation to real data, determine change points and peaks, identify the stabilization time of the epidemic, and provide predictions [2, 3]. The models should allow to visualize the possible twists and turns of the epidemics. In the case of the appearance of new waves, the model should register and illustrate such events. The pandemic research has activated scientists from various disciplines: biologists, epidemiologists, physicists, and of course also mathematicians [4, 5].

In this study, the cumulative number of individuals with COVID-19 during the pandemic period (since December 31, 2019) is modelled. The proposed technique allows fitting a large spectrum of statistical curves related to the growth process on a day-per-day basis. Among the curves used are the family of the sigmoidal growth functions and cumulative distribution functions. The computer program that realizes the applied numerical approach is also presented. The program allows to

realize various epidemiological models for any country, since the corresponding data are freely available. Here we constructed the models using a double Gompertz function and cumulative distribution of the probabilities [6-8]. Thus, the fitted model is a composition of three components. In this paper, the data were analyzed for four countries (Poland, Italy, Canada, USA), which have different number of the cases and dynamics of the epidemics of COVID-19.

2. Material and Methods

The data for this study were obtained from the web page of the organization "Our World in Data" [9]. The data contain daily values related to COVID-19 infections in most countries of the world. In this work, only daily cumulative cases and daily new cases of infection were used. The considered data are available for the period from December 31, 2019 to the current date. In this study, the data used were until September 10, 2020; 255 consecutive days in total. The analysis was done for four countries (Poland, Italy, Canada and the USA). Since we provide the computer program used in our analysis, it is relatively easy to perform similar calculations for any chosen country, and any chosen time period with the available data.

In our analysis, we applied the statistical software *R* [10], using the package *minpack.lm*. From this package, we used the function *nlsLM*. This function realizes standard least squares estimations of the parameters of a nonlinear model. The applied numerical method incorporates the Levenberg-Marquardt fitting algorithm [11]. Thus, for the given data on cumulative cases (CC) we fitted for them some non-linear functions along the considered sequence of days. Here we propose to use more than one function to represent fluctuations in the cumulative counts. In this work and for the considered countries we used the combination of three non-linear functions. The decision on the number of functions used and their type should be analyzed for each individual country. The criterion to choose a specific set of non-linear functions is the accuracy of the fitted models. The accuracy can be measured by various estimations, including the Akaike information criterion.

For the cumulative cases data for Poland and Italy we used two Gompertz functions and one normal distribution function. The normal function was realized using the function *pnorm* (in *R*, [10]), which evaluates the cumulative distribution function (cdf) of the normal distribution [7, 12]. In the case of Canada and the USA, we used two Gompertz functions and one gamma distribution function. For these countries we realized the cdf of the gamma distribution [7, 12].

For the four countries considered, to represent the cumulative cases, we applied the Gompertz function given by the formula with three parameters and two exponential functions.

$$y(t) = ae^{-be^{-ct}}$$

In the above equation, the parameters a, b, and c have the following interpretation; a-is an asymptotic value and is the limit of the function when t tends to infinity, b-is the displacement on the t-axis, c-represents the growth rate. In our analysis, time variable t is measured in days [10].

In summary, we used the following two approaches, where DAYD represents days (DAYD = t, time, CC-cumulative cases):

For Poland and Italy, we have two Gompertz functions and the normal distribution function (cdf) [10]. pnorm is the R function that calculates the cdf in the case of the normal distribution.

CC = M* exp(-A*exp(-B*DAYD)) + K*exp(-C*exp(-D*DAYD)) + L*pnorm(DAYD, E, F).

For Canada and the USA, we have two Gompertz functions and the gamma distribution (cdf) [10]. pgamma is the R function that calculates the cdf in the case of the gamma distribution.

CC = M* exp(-A*exp(-B*DAYD)) + K*exp(-C*exp(-D*DAYD)) + L*pgamma(DAYD, E, F).

The above notation is also used to identify the curves in Table 1 and Figure 1. The same information is provided in the Notes under Table 1.

Table 1 The coefficients of the fitted models for Poland, Italy, Canada, and the USA.

A-M	Poland	Italy	Canada	USA
A	9.294e+00	1.626e+02	7.013e+02	6.812e+02
B	5.676e-02	5.529e-02	2.792e-02	3.175e-02
C	4.339e+01	4.087e+01	1.124e+02	7.325e+00
D	4.068e-02	8.647e-03	4.131e-02	-1.015e-01
E	1.673e+02	8.320e+01	4.449e+02	2.476e+01
F	2.012e+01	4.647e+00	5.817e+00	2.042e-01
K	2.331e+04	3.661e+06	1.114e+05	-2.469e-03
L	3.617e+04	2.191e+04	-1.471e+03	1.931e+06
M	1.810e+04	2.188e+05	4.095e+04	5.424e+06

Notes-Models description. A-M the coefficients in the models: M* exp(-A*exp(-B*DAYD)) + K*exp(-C*exp(-D*DAYD)) + L*pnorm(DAYD, E, F). For Canada and the USA: L*pgamma(DAYD, E, F). DAYD-the number of days (1-255, from December 31, 2019 to September 10, 2020).

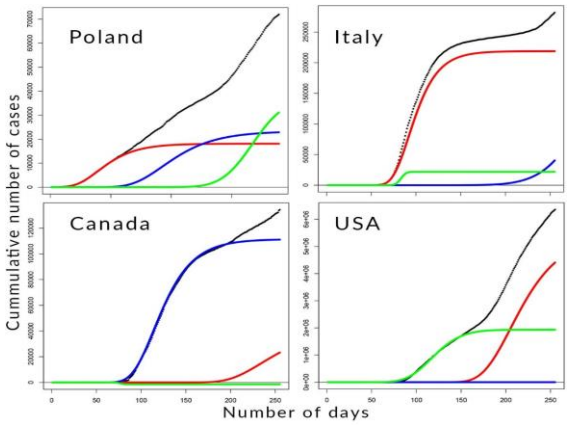

Figure 1 Three components of the fitted growth functions in Poland, Italy, Canada, and the USA. December 31, 2019-September 10, 2020. Black lines-original daily cumulative cases. Red (M) and blue (K) lines show the Gompertz function. Green lines identify the cdf functions. M and K identify the part of the constructed models (see the Note-Models description).

3. Results

The results are presented in the form of the fitted statistical models. The parameters of the functions are estimated. For the given health data (here, these are the counts of the cumulative cases of COVID-19 disease) a parametric non-linear function is constructed. Figure 2 gives the results for the considered countries, here Poland, Italy, Canada, and the USA. The figure has four panels

and shows the cumulative cases (black dots), fitted functions (red line), and daily new cases (multiplied by 20 to scale and shown in blue). The panels correspond to the indicated country. In the illustrated cases the fitted curves (red) overlap the original data points (black).

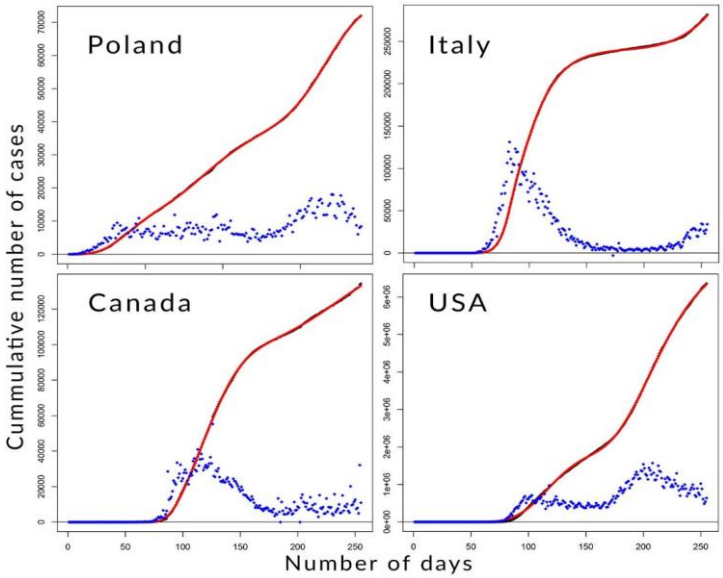

Figure 2 Non-linear growth curves to represent the number of cumulative cases in Poland, Italy, Canada, and the USA. December 31, 2019-September 10, 2020. Fitted line-red (the line overlaps original data-black). Blue colour shows the corresponding daily counts scaled by 20.

Table 1 presents the coefficients of the fitted functions. Using these values, we are able to reconstruct the function and describe the dynamics of the disease. We have analytical functions of the numbers of cumulative cases along time (days). This also can be used to predict values.

Figure 1 illustrates the original data (cumulative cases) and shows three components of the fitted individual functions. The summation of theses three components (two Gompertz functions and one cdf function) results in one growth curve shown on Figure 2. The original data are shown as black dots. The composite curves are identified as follows: Gompertz function (M-red, K-blue), cdf function (L-green).

A main purpose of this paper is to present some methodologies including the growth models. As each day new data are coming the analysis used to produce Figure 2 was repeated for the period December 31, 2019-January 30, 2021. The presented figure (Figure 3) illustrates the cumulative counts for the updated period. In a similar way Figure 4 is an updated until May 17, 2021.

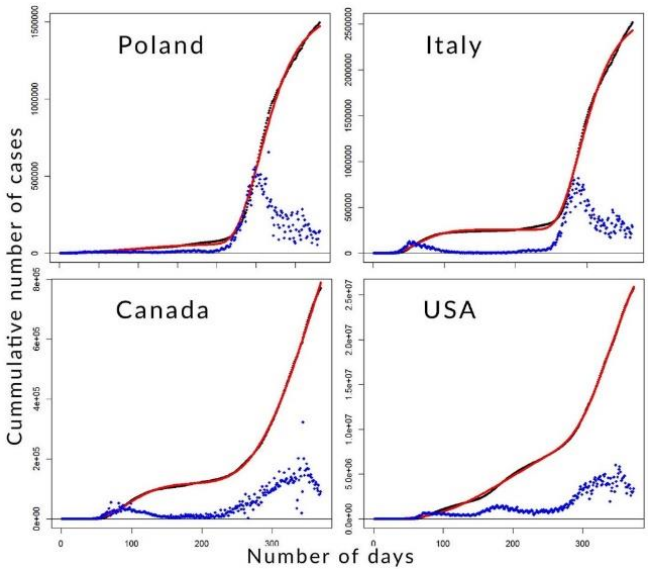

Figure 3 Non-linear growth curves to represent the number of cumulative cases in Poland, Italy, Canada, and the USA. December 31, 2019-January 30, 2021. Blue colour shows the corresponding daily counts scaled by 20.

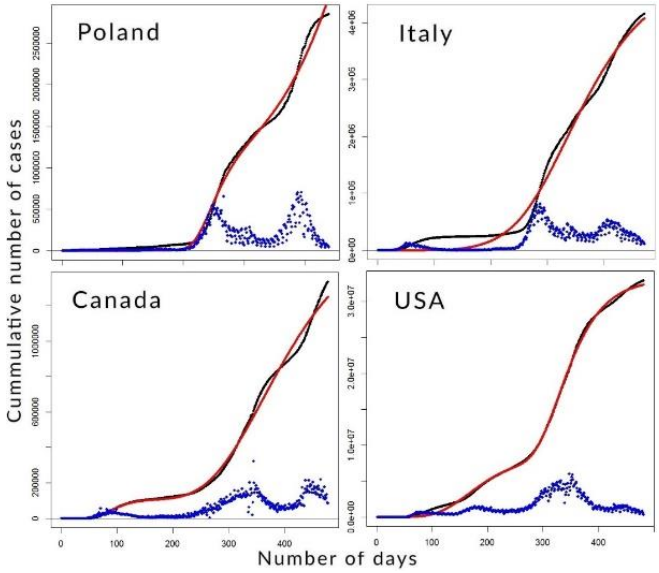

Figure 4 Non-linear growth curves to represent the number of cumulative cases in Poland, Italy, Canada, and the USA. December 31, 2019-May 17, 2021. Blue colour shows the corresponding daily counts scaled by 20.

Below in Appendix I is presented the program in the *R* statistical software. The program, for a given set of initial parameters, estimates the non-linear growth curves. The program also draws the original data and values of the fitted functions. It performs the calculations and does Figure 2.

4. Discussion

The Gompertz function and the cdf's of two probability distributions (gamma and normal) were used to fit the COVID-19 spread dynamics [10].

In this work, two kinds of approaches were proposed to represent the cumulative number of infected persons. One approach is to use a formula, which describes the parametric function of the growth phenomenon. These parameters are determined by the least square algorithm. Here we used the Gompertz function with three parameters a, b, and c. Another such example is the sigmoidal mathematical model proposed by Boltzmann in 1879 [13]. His method is based on the sigmoidal logistic equation

$$y(t) = \frac{1}{1 + e^{-t}}$$

Boltzmann model can be expressed as the following parametric function

$$y(t) = a + \frac{b - a}{1 + e^{\frac{c-t}{d}}}$$

where the usual starting values for its two parameters are as follows a=maximum of (y), b=minimum of (y). Practically, such functions are usually already programmed and included in the statistical software. We can find, in the *R* software, the function (*richards*) that generates the values of the Richards growth law and also the function *gompertz*, which produce values of the Gompertz growth [10]. Thus, in both cases we can just call these functions in the constructed models.

The Richards growth function can be described and represented by the following formula

$$y(t) = a * \left[1 + b * \exp\left\{1 + \mathrm{b} + \frac{c}{a}(1 + b)^{1+\frac{1}{b}} * (d - t)\right\}\right]^{\frac{-1}{b}}$$

In this case, we also have four parameters (a, b, c, and d) as in the Gompertz function to determine to approximate the cumulative cases [10, 14].

We can also use the packages designed specially to build the growth response. One such example is the package *growthcurver* [15].

Another approach applied in this work was to use the shapes of the cdf's of the gamma and normal distribution function. We can also use other distributions suitable in this context. One good candidate is the Weibull distribution. In the *R* software, we have the cdf function *pweibull*, which can be used in a similar way as in the case of the gamma and normal distribution [7, 16].

We observed that our models with three components result in a good approximation. The simplest fit and a lower number of applied functions is the desired approach. As we can see, we have many different options. We can use a few functions of the same type or a mixture of various types.

In general, we are able to represent the cumulative number of infected persons in the form of algebraic functions. We can use the obtained formulae to construct predictions, analyze the spread properties, and estimate changes along time. In the examples considered, we included the data with zero cases, in the early period. The fitted curves are flat for beginning unit days (no cases) until the first case of the COVID-19 disease.

Here, we investigated and represented the spread dynamics of the COVID-19 disease in four countries. The study was based on the use of a few mathematical models. Using the freely available

statistical software *R*, it is relatively easy to conduct a similar analysis for other countries and more recent data. The data can be analyzed in different time intervals.

There are many approaches which can be applied to validate the time-series method [17-21]. The set of existing articles that explore COVID-19 forecasts or curve exploration is large [20-22]. This study proposed to use known methods (growth models, CDF) as alone or in a combination with other models.

In a study recently published it is shown that an exponential decay model applied to the weighted and averaged growth rates appears to be better than Gompertz's model for modeling the number of cases of the COVID-19 [23].

Additional Materials

The following additional materials are uploaded at the page of this paper.

1. Appendix I

Author Contributions

The author did all the research work of this study.

Competing Interests

The author has declared that no competing interests exist.

References

1. WHO coronavirus (COVID-19) dashboard. World Health Organization; 2020 [cited 2020 October 16]. Available from: https://covid19.who.int/?gclid=EAIaIQobChMI2-vft9iY7AIVxZyzCh3WJgugEAAYASAAEgIOf_D_BwE.
2. Ciufolini I, Paolozzi A. Mathematical prediction of the time evolution of the COVID-9 pandemic in Italy by a Gauss error function and Monte Carlo simulations. Eur Phys J Plus. 2020; 135: 355.
3. Zhang XL, Ma RJ, Wang L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos Solitons Fractals. 2020; 135: 109829.
4. Copiello S, Grillenzoni C. The spread of 2019-nCoV in China was primarily driven by population density. Comment on "Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China" by Zhu et al. Sci Total Environ. 2020; 744: 141028.
5. Davies NG, Klepac P, Liu Y, Prem K, Jit M, CMMID COVID-19 working group, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med. 2020; 26: 1205-1211.
6. Gompertz B. XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c. Philos Trans R Soc. 1825; 115: 513-583.
7. Feller W. An introduction to probability theory and its applications. 3rd ed. New York: Wiley; 1968.
8. Monti KL. Folded empirical distribution function curves–mountain plots. Am Stat. 1995; 49: 342-345.

9. Coronavirus pandemic (COVID-19)-the data. Our World in Data; [cited 2021 June 3]. Available from: https://ourworldindata.org/coronavirus-data.
10. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
11. Marquardt D. An algorithm for least-squares estimation of nonlinear parameters. SIAM J Appl Math. 1963; 11: 431-441.
12. Weisstein EW. Normal distribution. From MathWorld--a Wolfram web resource. Wolfram MathWorld; [cited 2020 October 16]. Available from: https://mathworld.wolfram.com/NormalDistribution.html.
13. Reséndiz-Muñoz J, Corona-Rivera MA, Fernández-Muñoz JL, Zapata-Torres M. Mathematical model of Boltzmann's sigmoidal equation applicable to the set-up of the RF-magnetron co-sputtering in thin films deposition of $Ba_xSr_{1-x}TiO_3$. Bull Mater Sci. 2017; 40: 1043-1047.
14. Richards FJ. A flexible growth function for empirical use. J Exp Bot. 1959; 10: 290-301.
15. Sprouffske K, Wagner A. Growthcurver: An R package for obtaining interpretable metrics from microbial growth curves. BMC Bioinform. 2016; 17: 1-4.
16. Weisstein EW. "Weibull distribution." From MathWorld--a Wolfram web resource. Wolfram MathWorld. Available from: https://mathworld.wolfram.com/WeibullDistribution.html.
17. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons. Int J Forecast. 1992; 8: 69-80.
18. Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev VA. A survey of forecast error measures. World Appl Sci J. 2013; 24: 171-176.
19. Lynch CJ, Gore R. Short-range forecasting of COVID-19 during early onset at county, health district, and state geographic levels using seven methods: Comparative forecasting study. J Med Internet Res. 2021; 23: e24925.
20. Lynch CJ, Gore R. Application of one-, three-, and seven-day forecasts during early onset on the COVID-19 epidemic dataset using moving average, autoregressive, autoregressive moving average, autoregressive integrated moving average, and naïve forecasting methods. Data Brief. 2021; 35: 106759.
21. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect Dis Model. 2020; 5: 256-263.
22. Singh RK, Rani M, Bhagavathula AS, Sah R, Rodriguez-Morales AJ, Kalita H, et al. Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model. JMIR Public Health Surveill. 2020; 6: e19115.
23. Bartolomeo N, Trerotoli P, Serio G. Short-term forecast in the early stage of the COVID-19 outbreak in Italy. Application of a weighted and cumulative average daily growth rate to an exponential decay model. Infect Dis Model. 2021; 6: 212-221.

Enjoy *AEER* by:

1. Submitting a manuscript
2. Joining in volunteer reviewer bank
3. Joining Editorial Board
4. Guest editing a special issue

For more details, please visit:
http://www.lidsen.com/journals/aeer