Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Applied Soft Computing 105 (2021) 107289

Contents lists available at ScienceDirect

Applied Soft Computing

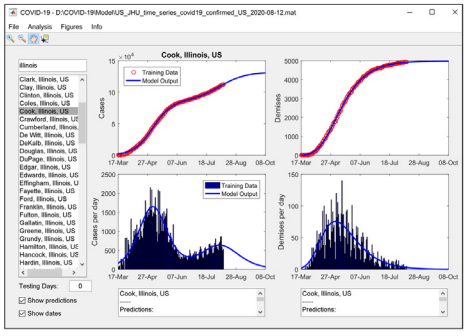
journal homepage: www.elsevier.com/locate/asoc

A computational tool for trend analysis and forecast of the COVID-19 pandemic

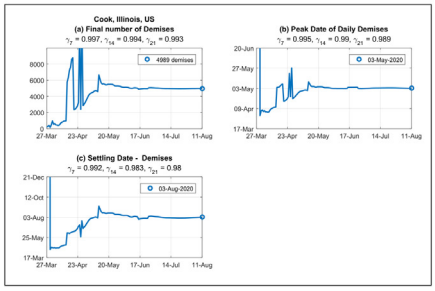
Henrique Mohallem Paiva [a,*], Rubens Junqueira Magalhães Afonso [b,c], Fabiana Mara Scarpelli de Lima Alvarenga Caldeira [a], Ester de Andrade Velasquez [a]

[a] Institute of Science and Technology (ICT), Federal University of Sao Paulo (UNIFESP), Rua Talim, 330, São José dos Campos, SP, Brazil
[b] Institute of Flight System Dynamics, Technical University of Munich (TUM), München, Bayern, 85748, Germany
[c] Department of Electronic Engineering, Aeronautical Institute of Technology (ITA), Praça Marechal Eduardo Gomes, 50, São José dos Campos, SP, Brazil

GRAPHICAL ABSTRACT



Graphical User Interface (GUI)



Trend Analysis

ARTICLE INFO

Article history:
Received 19 August 2020
Received in revised form 22 February 2021
Accepted 5 March 2021
Available online 10 March 2021

Keywords:
COVID-19
Epidemiology
Mathematical modeling
Trend analysis
Forecast
Numerical optimization
Sequential quadratic programming (SQP)

ABSTRACT

Purpose: This paper proposes a methodology and a computational tool to study the COVID-19 pandemic throughout the world and to perform a trend analysis to assess its local dynamics.
Methods: Mathematical functions are employed to describe the number of cases and demises in each region and to predict their final numbers, as well as the dates of maximum daily occurrences and the local stabilization date. The model parameters are calibrated using a computational methodology for numerical optimization. Trend analyses are run, allowing to assess the effects of public policies. Easy to interpret metrics over the quality of the fitted curves are provided. Country-wise data from the European Centre for Disease Prevention and Control (ECDC) concerning the daily number of cases and demises around the world are used, as well as detailed data from Johns Hopkins University and from the Brasil.io project describing individually the occurrences in United States counties and in Brazilian states and cities, respectively. U. S. and Brazil were chosen for a more detailed analysis because they are the current focus of the pandemic.
Results: Illustrative results for different countries, U. S. counties and Brazilian states and cities are presented and discussed.
Conclusion: The main contributions of this work lie in (i) a straightforward model of the curves to represent the data, which allows automation of the process without requiring interventions from experts; (ii) an innovative approach for trend analysis, whose results provide important information to support authorities in their decision-making process; and (iii) the developed computational tool, which is freely available and allows the user to quickly update the COVID-19 analyses and forecasts

* Corresponding author.
E-mail address: hmpaiva@unifesp.br (H.M. Paiva).

https://doi.org/10.1016/j.asoc.2021.107289

for any country, United States county or Brazilian state or city present in the periodic reports from the authorities.

1. Introduction

On December 2019, a series of pneumonia cases of unknown cause emerged in Wuhan, China, with clinical presentations greatly resembling viral pneumonia [1]. The Chinese authorities identified a new type of coronavirus (novel coronavirus, named 2019-nCoV), which was isolated on 7 January 2020 [2]. Coronaviruses are a family of viruses that can cause respiratory, hepatic, and neurological diseases in humans and animals [3].

Initially, these infections were thought to result from zoonotic (animal-to-human) transmission. However, an exponential growth of case incidence, with many cases detected in other parts of the world, showed a strong evidence of human-to-human secondary transmission [4]. On March 2020, the coronavirus disease (COVID-19) was declared a public health emergency of international concern by the World Health Organization [5]. A worrying aspect in relation to the disease is the fact that it is highly contagious, spreading very quickly and causing overcrowding in the health system [6].

Since January 2020, many studies have been done to assess the transmission potential of 2019-nCoV nationally and internationally, as well as to forecast its spread. Quarantine measures have been implemented worldwide, as international travel has helped to spread the virus to other parts of the world [6,7]. At the time of this writing, 57.8 million people worldwide were infected with COVID-19 and 1.3 million passed away [8].

Many efforts have been done to determine an effective treatment and to develop a vaccine [9–11]. Currently, the results of many studies suggest that early detection, hand washing, self-isolation and household quarantine are effective to mitigate this pandemic [7].

Similar situations, although in smaller scale, occurred during the Influenza A epidemic in 2009 and during the Middle East Respiratory Syndrome coronavirus (MERS-CoV) epidemic in 2012. The Influenza A virus appeared in April 2009 [12] and caused a pandemic with more than 280,000 deaths worldwide [13]. The MERS epidemic emerged in Saudi Arabia in 2012 [14] and caused thousands of infections in dozens of countries worldwide. This virus, also belonging to the coronavirus family, has a high fatality rate [15]. Under these scenarios, Dugas et al. [16] created a forecasting model for Influenza A and Kim et al. [14] formulated a forecasting model for MERS transmission dynamics and estimated transmission rates, considering several categories of patients and transmission rates. In fact, mathematical models have been widely used to study the transmission dynamics of infectious diseases, enabling the understanding of the disease spread and the optimization of disease control [17].

Forecasting models are used to predict future behavior as a function of past data. This is a widely used method in the implementation of epidemic mathematical models, since it is necessary to know the past behavior of a disease to understand how it will evolve in the future. Accurate forecasts of disease activity could allow for better preparation, such as public health surveillance, development and use of medical countermeasures, and hospital resource management [18].

A similar approach is the concept of trend analysis, which allows predicting future behavior with accuracy, especially in the short run. A trend is a change over time exhibited by a random variable [19]; trend analyses provide direction to a trend from past behavior, allowing predicting future data. For better effectiveness, the predictions should be updated periodically, as soon as new data are available.

The technique of trend analysis is widely used in several areas of science, such as finances [20,21] and meteorology [19,22]. In the context of health systems, trend analysis was used by Zhao et al. [23], to analyze malignant mesotheliomas in China, aiming to provide data for its prevention and control; by Soares et al. [24], to predict the testicular cancer mortality in Brazil; by Zahmatkesh et al. [25], to forecast the occurrences of breast cancer in Iran; by Mousavizadeh et al. [26], to forecast multiple sclerosis in a region of Iran; and by Yuan et al. [27], to analyze and predict the cases of type 2 diabetes in East Asia.

Modeling and prediction of the dynamics of the COVID-19 pandemic is a subject of great interest. Therefore, a myriad of papers on this theme have been published over the last months. For this purpose, some research groups extended previous epidemiological models to describe the COVID-19 pandemic: Lin et al. [28] created a conceptual model for the COVID-19 outbreak in Wuhan, China, using components from the 1918 influenza pandemic in London, while Paiva et al. [29] proposed a dynamic model to describe the COVID-19 pandemic, based on a model previously developed for the MERS epidemic. Different modeling approaches have been exploited, such as compartment models [30], time series analysis [31], artificial intelligence [32,33], and regression-based models [34,35]. This list is far from being exhaustive. For a detailed survey on different modeling approaches in this context, the reader is referred to review papers such as [36] and [37].

It is important to note that the behavior of the pandemic may vary greatly in the different regions of the world, due to characteristics such as different social habits (higher or lower physical interaction between citizens), capacity of the local health system, different governmental actions, and so on. Therefore, the parameters of a mathematical model need to be tailored to the region where the disease behavior is being studied. Furthermore, even in the same region, the conditions may vary very quickly, in a matter of weeks or even days (for instance, following the decree or release of a lockdown, or the saturation of the available intensive care unit vacancies in the hospitals); thus, the model parameters would need to be updated very often, usually by an expert. However, these analyses might take time and require dedicated work from highly qualified personnel, thus decreasing their availability. It is natural to expect that such analyses are run periodically at the country level, but the same may not be a reality locally at every municipality. Therefore, in this scenario, it is useful to have a computational tool to perform a quick and automatic analysis and forecast of the disease conditions in any region, following the periodic updates published by the authorities. This is the purpose of the present paper.

The methodology proposed here employs mathematical functions to model the behavior of the pandemic. A numeric optimization algorithm is used to calibrate such models, in an automatic process that does not require intervention from experts. An original trend analysis technique is proposed, allowing determining the effects of public policies. Illustrative results at different territorial levels (country, state, and city) are presented and discussed. A computational tool was developed and is available online, allowing to process data from different countries and subnational data from the United States and Brazil.

When compared to other modeling approaches from the literature, the main advantage of our model lies in the automatic

calibration process, which allows the analysis in different regions of the world, being especially useful for regions where experts are not available. The automatic analysis is also useful to update the results as soon as new data becomes available. Furthermore, the automation also allows the proposed trend analysis, which would not be feasible manually because it requires an impractically large number of parameter estimations with different amounts of data.

In the remainder of the paper, we first present the fundamental mathematical function adopted here, which is an asymmetric sigmoid, as well as its tuning parameters and the most relevant epidemic characteristics that can be inferred from the curve. The procedure to fit the data through optimization is then discussed. Subsequently, we discuss criteria to define the complexity of the model, i.e., whether a symmetric sigmoid is enough to describe the data adequately or an asymmetric one is necessary and also the number of sigmoids that should be used for fitting a set of data for a locality. Afterwards, we discuss when it can be considered that the convergence to the final value has happened. Statistical tools to evaluate the quality of the fit are presented in the sequence. After that, results are presented for several localities illustrating the properties of the proposal and the usage of the software. Then an in-depth discussion of the chosen examples is carried out and finally conclusions are drawn.

2. Methods

This section describes the methods adopted in this study. Section 2.1 discusses the mathematical foundations employed to formulate the model describing the behavior of the pandemic. Section 2.2 explains how the model parameters are estimated. Section 2.3 extends the formulation to analyze a repeated behavior, characterized by multiple occurrences of the fundamental function. Sections 2.4 and 2.5 describe the criteria to evaluate the suitability of the estimated parameters, by quantifying the accuracy of the model output compared to historical data and the convergence properties of the estimations. Sections 2.6 and 2.7 explain the methodologies to select the complexity of the model and to quantify the associated uncertainties. Section 2.8 presents a summary. Finally, the computer implementation is described in Section 2.9.

2.1. Mathematical formulation of the curve to describe the data

In the present paper, the fundamental curve that is used to describe the historical data is an asymmetric sigmoid, i.e., letting the independent variable be t, then the dependent variable is given as a function $f: \mathbb{R} \mapsto \mathbb{R}^+$ [38]:

$$f(t) = \frac{A}{\left(1 + \nu e^{-\frac{(t-t_p)}{\delta}}\right)^{\frac{1}{\nu}}} \tag{1}$$

with the parameters $A, t_p \in \mathbb{R}^+ \cup \{0\}$ and $\nu, \delta \in \mathbb{R}^+$. A is the final number of occurrences; t_p is the day with maximum daily occurrences; ν is a parameter defining how asymmetric is the function; and δ is a parameter associated to how fast the convergence of the function to its final value A is. The meaning of these four parameters will become clearer in the forthcoming discussion.

In the present work, the independent variable t is the time in days, whereas the dependent variable is either the cumulative number of individuals that were positively tested for SARS-CoV-2 or the cumulative number of individuals deceased with the disease as the cause.

Notice that

$$\lim_{t\to\infty} f(t) = A \tag{2}$$

i.e., the modeling of the cumulative number of cases/demises by (1) implies convergence to a final value A. However, the convergence is asymptotic, therefore it is interesting to know when a certain threshold of the final number of infected/deceased has been reached. For that purpose, let a time instant τ_α be such that a particular value $f(\tau_\alpha)$ is reached:

$$f(\tau_\alpha) = \alpha A \tag{3}$$

where the parameter $\alpha \in]0, 1[$. Then, by replacing (1) for $f(\tau_\alpha)$ in (3), one may solve to find:

$$\tau_\alpha = t_p - \delta \ln\left\{\frac{1}{\nu}\left[\left(\frac{1}{\alpha}\right)^{\nu} - 1\right]\right\} \tag{4}$$

Therefore, from (4) one can determine the (finite) instant when a certain proportion of the final number of cases/demises is reached, which is a useful figure to evaluate whether the contamination can be considered over or not. In this paper, the settling date of the contamination is adopted as $\tau_{0.98}$, corresponding to the day where the number of occurrences reaches 98% of its final value. The settling ratio of 98% is a standard value used in the analysis of dynamic systems [39].

The rate at which the number of infections/demises grows can be calculated by differentiation of (1) with respect to the independent variable t, which yields

$$\frac{df(t)}{dt} = \frac{A}{\delta}\frac{g(t)}{(1 + \nu g(t))^{\frac{\nu+1}{\nu}}} \tag{5}$$

where

$$g(t) = e^{-\frac{(t-t_p)}{\delta}} \tag{6}$$

As a matter of fact, the value of (5) in a particular day t is an important indicator for healthcare infrastructure decision-making concerning the number of infected individuals, as a higher value indicates that the upcoming period might stress the healthcare infrastructure, whereas a comparatively lower value points that the number of new cases might be accommodated with the existing infrastructure. By analyzing the number of individuals that are cured each day and discharged from the facilities and comparing it with the rate of newly infected individuals, if the first is greater than the latter, than the capacity of the facilities is enough to treat the ill and they will not be endangered by lack of proper treatment.

Differentiating (5) with respect to t yields

$$\frac{d^2 f(t)}{dt^2} = \frac{A}{\delta^2} g(t) \frac{g(t) - 1}{(1 + \nu g(t))^{\frac{2\nu+1}{\nu}}} \tag{7}$$

A sign change in (7) occurs when the term $(g(t) - 1)$ crosses zero, as the remaining terms are all positive for any $t \in \mathbb{R}$. Therefore, there is a single inflection in the curve (5) at $t = t_p$. On the other hand, since $\frac{d^2 f(t)}{dt^2} > 0$ for $t < t_p$ and $\frac{d^2 f(t)}{dt^2} < 0$ for $t > t_p$, this point corresponds to the maximum rate, i.e., the daily number of either infected or deceased individuals. Replacing $t = t_p$ in (1) yields

$$f(t_p) = \frac{A}{(1 + \nu)^{\frac{1}{\nu}}} \tag{8}$$

Notice from (8) that $\nu = 1$ entails $f(t_p) = 0.5A$, i.e., the sigmoid curve crosses half of the final value at $t = t_p$. This is deemed a symmetric sigmoid. For the sake of understanding, consider two other illustrative possible values of ν:

(a) for $\nu = 2$, from (8), $f(t_p) = A/\sqrt{3} \approx 0.58A$, that is, the inflection happens at a later stage, when roughly 58% of the final value has been reached;

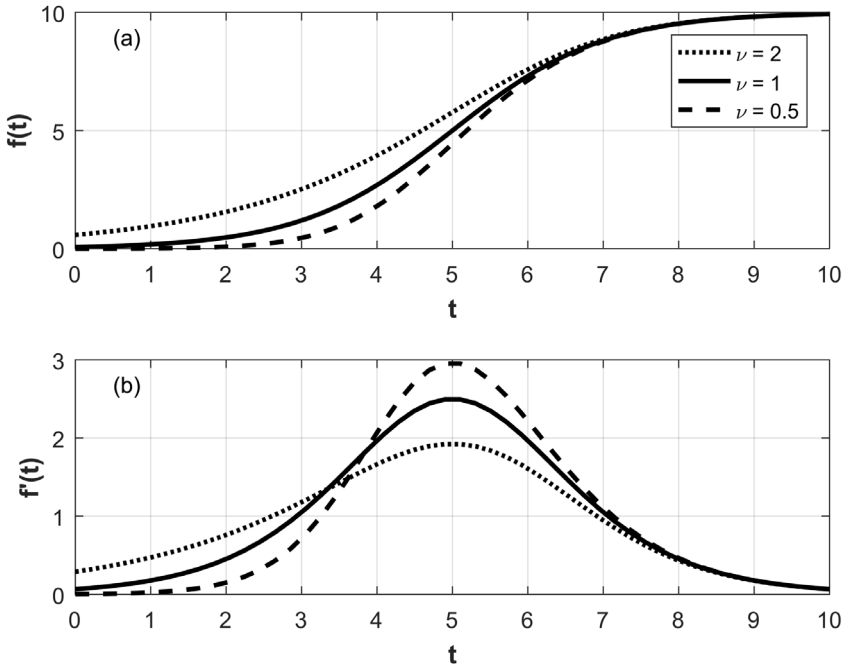


Fig. 1. (a) Sigmoid curves and (b) their derivatives for different values of the parameter ν. The remaining parameters are $A = 10$, $t_p = 5$, and $\delta = 1$.

(b) for $\nu = 0.5$, from (8), $f(t_p) = 4A/9 \approx 0.44A$, in other words, the inflection happens at an earlier stage, when approximately only 44% of the final value has been reached.

It is clear from these examples and from (1) that the value of ν controls the degree of asymmetry in the sigmoid curve, with $\nu = 1$ representing a symmetric curve about the $t = t_p$ vertical straight-line. This is illustrated in Fig. 1(a), where (1) is shown for three values of ν whereas Fig. 1(b) shows (5), i.e., the rate. It is interesting to remark that the value of ν impacts the symmetry of the derivative, with $\nu = 1$ representing a Gaussian curve, with acceleration and deceleration phases occurring at the same rate. When $\nu < 1$, the deceleration phase of the sigmoid is slower than the acceleration phase; when $\nu > 1$, the opposite occurs.

In view of their capability of representing processes with asymmetric acceleration and deceleration phases, asymmetric sigmoid curves are interesting to represent the data of a pandemic. Many factors can contribute to the asymmetry between acceleration and deceleration phases besides the very nature of the disease spread, such as the introduction of policies by health authorities in order to slow down the spread, e.g., reduced social contact. Therefore, this extra degree of freedom brought by the asymmetric sigmoid curve is useful to better represent the data. Moreover, the added complexity with regard to a symmetric curve is due only to the necessity of estimating a single additional parameter, namely ν.

In our context, there are three main sigmoid parameters of interest, which are described in Table 1.

The next subsection presents the algorithm used to estimate the parameters A, ν, δ, and t_p based on measured data from either the number of newly infected individuals per day or the number of deceased per day.

2.2. Parameter estimation

The parameters A, ν, δ, and t_p are estimated based on the solution of a constrained optimization problem, in which the Integral Time Square Error (ITSE) [39] is minimized, where the error is the difference between the value of f(t) output by (1) and the corresponding data y(t) obtained from the authorities at the same day. We consider a time window for $t \in \{0, 1, \ldots, t_{end}\}$ for which the data y(t) are available at each day. There is a small abuse of notation by restricting the real-valued variable t to assume only integer values coinciding with the number of the day,

Let the vector of parameters to be estimated be defined as

$$\Theta = \begin{bmatrix} A & \nu & \delta & t_p \end{bmatrix}^T \tag{9}$$

where the symbol $\bullet^T$ indicates the transpose of a vector $\bullet$. The optimal value of the vector Θ is given as

$$\Theta^* = \underset{\{A \geq 0,\ \nu > 0, \delta > 0, t_p \geq 0\}}{\operatorname{argmin}} \sum_{t=0}^{t_{end}} t\,[y(t) - f(t, \Theta)]^2 \tag{10}$$

where the argument Θ was explicitly included in $f(t, \Theta)$ to emphasize that the parameters may be varied during the optimization process. Note that, for optimization purposes, strict inequalities cannot be implemented, therefore for the constraints $\nu > 0$ and $\delta > 0$, an arbitrary small positive real number $\mu > 0$ is chosen and the constraints are approximated as $\nu \geq \mu$ and $\delta \geq \mu$. After the optimization problem is solved to yield Θ^*, the optimal values of A^*, ν^*, δ^*, t_p^* are fixed values used to build the curve.

The function $f(t, \Theta)$ is nonlinear in the parameters Θ, and the cost function exacerbates that further, rendering the optimization problem nonlinear. Moreover, the inequality constraints introduce additional difficulty, rendering the analytical solution of the optimization problem impractical. Therefore, numerical methods must be used.

One class of methods that are suitable for nonlinear constrained optimization is the so-called Sequential Quadratic Programming (SQP) [40–42]. SQP iteratively approximates the general nonlinear cost function in (10) by a quadratic one, and the constraints by linear ones, which entails a Quadratic Programming (QP) problem. QPs can be solved to global optimality in finite time, therefore each iteration of the SQP method takes finite time. The solution of the underlying QP approximation is then used to build a next iterate, for which another QP is solved, therefore the name Sequential Quadratic Programming. SQP presents

Table 1
Main sigmoid parameters of interest.

Parameter	Eqs.	Description
A	(1)(2)	Final number of occurrences.
t_p	(1)	Date when the maximum number of daily occurrences is achieved.
$\tau_{0.98}$	(3)(4)	Settling date, defined as the date when the number of occurrences reaches 98% of its final value.

good convergence properties, converging quadratically to the optimal solution when the active set does not change [43]. The implementation of SQP that is used in the present work is that of the function **fmincon** [44], from the Optimization Toolbox ™ of MATLAB®.

In order to take into account the dependence between cases and demises in the same region, we initially estimate the function describing the number of cases and then impose two additional constraints for the function representing the behavior of the demises. Considering that the demises will necessarily occur after the infections, the two additional constraints impose that both the date of maximum daily occurrences and the settling date for the demises must occur after the corresponding dates in the function describing the number of cases.

2.3. Multiple sigmoids

A second wave of spread has not been discarded. On the contrary, researchers argue that lifting the social distance measures might indeed lead to a retake in the infections [45–48].

In order to describe the occurrence of multiple epidemiological waves, we propose to employ a sum of sigmoids. For this purpose, let N_s be the adopted number of sigmoids. Eq. (1) is then generalized to

$$f(t) = \sum_{i=1}^{N_s} f_i(t) \tag{11}$$

where

$$f_i(t) = \frac{A_i}{\left(1 + \nu_i e^{-\frac{(t-t_{p,i})}{\delta_i}}\right)^{\frac{1}{\nu_i}}}, \quad i = 1, 2, \ldots, N_s \tag{12}$$

Similarly, the vector of parameters Θ, originally given by (9), is generalized to a column vector with $4N_s$ parameters defined as:

$$\Theta = \begin{bmatrix} \Theta_1^T & \Theta_2^T & \ldots & \Theta_{N_s}^T \end{bmatrix}^T \tag{13}$$

where

$$\Theta_i = \begin{bmatrix} A_i & \nu_i & \delta_i & t_{p,i} \end{bmatrix}^T, \quad i = 1, 2, \ldots, N_s \tag{14}$$

With these extended definitions, Eq. (10) can still be used to estimate the value of Θ^* by considering the inequalities applied to each A_i, ν_i, δ_i and $t_{p,i}$, $i = 1, 2, \ldots, N_s$.

It is important to establish the number of sigmoids N_s. For this purpose, an evaluation of the number of switches between deceleration and acceleration phases is performed. The rationale behind this assessment is: each sigmoid results in a single acceleration and a single deceleration phases, with a clear switching point between them, as discussed in Section 2.1. Therefore, the number of sigmoids can be estimated by counting the amount of switches from a deceleration to an acceleration phase. However, this counting requires careful consideration, as one is dealing with real noisy data. More so, recall that for identifying acceleration/deceleration the second derivative of the cumulative number of either infected or deceased individuals has to be considered. As it is well known, differentiation is prone to increase the effect of noise in the measurements [39]. Therefore, to mitigate the effect of noise in increasing artificially the amount of switches, a common approach is to consider a deadzone [39] in the difference between the acceleration and deceleration.

Let S be the set of switching instants from a deceleration to an acceleration phase. Then, for each $t = 0, 1, \ldots, t_{end} - 1$, the following logic is used to implement an identification of switches with a deadzone:

$$\begin{aligned} &\frac{d^2f}{dt^2}(t) \le -\epsilon \text{ and } \frac{d^2f}{dt^2}(t+1) \ge \epsilon \Rightarrow \; t \in S \\ &\text{otherwise, } t \notin S \end{aligned} \tag{15}$$

where the parameter ϵ can be adjusted to provide a compromise between noise and detection sensitivity. In the present work, the value was set to $\epsilon = 3 \cdot 10^{-5}$ persons/day ^{2}.

Thus, the number of sigmoids is given by the cardinality of S, summed with 1.

$$N_s = |S| + 1 \tag{16}$$

The value of 1 refers to the first sigmoid.

Recall, from Table 1, that there are three parameters of interest. The final number of occurrences may be obtained as:

$$A = \sum_{i=1}^{N_s} A_i \tag{17}$$

On the other hand, when a sum of sigmoids is used, there are no analytical expressions to determine the other two parameters of interest, i.e., the date of maximum number of daily occurrences t_p and the settling date $\tau_{0.98}$. In this case, a numerical search algorithm has to be used to find each of these parameters.

The optimization problems to determine these parameters can be posed as follows:

$$t_p = \underset{t \ge 0}{\operatorname{argmax}} \frac{df}{dt}(t) \tag{18}$$

$$\tau_{0.98} = \underset{t \ge 0}{\operatorname{argmin}} [f(t) - 0.98A]^2 \tag{19}$$

These two optimization problems are solved using the Nelder–Mead algorithm [49]. It should be noted that each problem has only one independent variable (time). Therefore, the search algorithm converges very quickly to the desired solution.

The rationale to select the use of one or multiple sigmoids will be explained in Section 2.6, which discusses the complexity of the model.

2.4. Criteria for statistical analysis of the matching between the fitted curve and the data

Two criteria are used to evaluate the degree of fidelity of the fitted curves to the data. The first is the so-called Root Mean Square Error (RMSE), defined as:

$$RMSE = \sqrt{MSE} \tag{20}$$

$$MSE = \sum_{t=0}^{t_{end}} \frac{[y(t) - f(t, \Theta^*)]^2}{t_{end} + 1} \tag{21}$$

From (20) the name of RMSE becomes clear, as it involves the square root of the mean of the squared error (MSE). Notice that, in (21), the values of the curve with the optimal parameters $f(t, \Theta^*)$ are used to calculate the error between the data and the value returned by the fitted curve. Moreover, the term $t_{end} + 1$ reflects the number of terms in the summation, as the index t starts at 0 and ends at t_{end}. The RMSE is used in statistical analysis to measure compactly the degree of fidelity between the fitted curve and the data. The lower the value of the RMSE, the better the fitted curve matches the data [50].

In this paper, a normalized version of the RMSE is used, obtained as:

$$\text{normalized RMSE} = \frac{\text{RMSE}}{A} \tag{22}$$

where A is the final number of occurrences, as defined in (1) and (17) for one and multiple sigmoids, respectively. This normalization is adopted to allow a fair comparison of the RMSE of different curves.

A second criterion to determine the quality of the representation of the data by the fitted curve generally applied in statistics is the squared correlation coefficient, which varies between 0 and 1, with the latter meaning that there exists a perfect linear functional relationship between the data and the fitted curve points, whereas the first means the opposite. First, let us define the covariance of the data as

$$\text{cov}\left[y(\bullet), f(\bullet, \Theta^*)\right] = \sum_{t=0}^{t_{end}} \frac{\left[y(t) - \mu_y\right]\left[f(t, \Theta^*) - \mu_f\right]}{t_{end}} \tag{23}$$

where μ_y and μ_f are the mean values of $y(t)$ and $f(t, \Theta^*)$, respectively, i.e.

$$\mu_{\blacksquare} = \sum_{t=0}^{t_{end}} \frac{\blacksquare}{t_{end} + 1} \tag{24}$$

in which the symbol $\blacksquare$ can be replaced by either of $y(t)$ and $f(t, \Theta^*)$, yielding μ_y and μ_f, respectively. Similarly, the variances of $y(t)$ and $f(t, \Theta^*)$ are

$$\text{var}\left[y(\bullet)\right] = \text{cov}\left[y(\bullet), y(\bullet)\right] = \sum_{t=0}^{t_{end}} \frac{\left[y(t) - \mu_y\right]^2}{t_{end}} \tag{25}$$

$$\text{var}\left[f\left(\bullet, \Theta^*\right)\right] = \text{cov}\left[f\left(\bullet, \Theta^*\right), f(\bullet, \Theta^*)\right] = \sum_{t=0}^{t_{end}} \frac{\left[f(t, \Theta^*) - \mu_f\right]^2}{t_{end}} \tag{26}$$

The squared correlation coefficient R^2 can then be determined from (23)–(26) as:

$$R^2 = \frac{\left\{\text{cov}\left[y(\bullet), f(\bullet, \Theta^*)\right]\right\}^2}{\text{var}\left[y(\bullet)\right]\text{var}\left[f(\bullet, \Theta^*)\right]} \tag{27}$$

2.5. Criteria for assessment of convergence of the sigmoid towards the final value

Additional criteria are defined to evaluate whether the data are enough to allow the convergence of the estimated values of the parameters Θ^*. This is carried out by fitting the sigmoid curves to the data for each possible value of $n \in \{10, 11, \ldots, t_{end} - 21, t_{end} - 20, \ldots, t_{end}\}$. Thus, instead of using all available data as in (10), windows of varying length are used; the minimum length of a window is adopted as 10 to ensure a minimum amount of data to calibrate the curve. Therefore, the sigmoid parameters Θ are estimated within different windows as

$$\Theta^*(n) = \underset{\left\{A_i \geq 0,\ \nu_i > 0,\ \delta_i > 0,\ t_{p,i} \geq 0\right\}}{\text{argmin}} \sum_{t=0}^{n} t\left[y(t) - f(t, \Theta)\right]^2 \tag{28}$$

where $i = 1, 2, \ldots, N_s$, depending on the number of sigmoids.

The main parameters in Table 1 are then determined from $\Theta^*(n)$ as follows:

- For a single sigmoid, $A^*(n)$ and $t_p^*(n)$ are directly extracted from $\Theta^*(n)$ in view of (9), whereas $\tau_\alpha^*(n)$ is calculated by (4) employing $t_p^*(n)$, $\delta^*(n)$ and $\nu^*(n)$ extracted from $\Theta^*(n)$ considering (9).
- For multiple sigmoids, (17)–(19) are used to determine $A^*(n)$, $t_p^*(n)$ and $\tau_{0.98}^*(n)$.

Then, the relative variation of the estimated values of these parameters is calculated for each time window and multiplied over the time window, composing indices to evaluate if the data are enough to asseverate the suitability of the sigmoid that was fitted. These indices are defined as

$$\gamma_k^{\blacktriangle} = \prod_{j=1}^{k} \min\left(\frac{\blacktriangle^*(t_{end} - j - 1)}{\blacktriangle^*(t_{end} - j)}, \frac{\blacktriangle^*(t_{end} - j)}{\blacktriangle^*(t_{end} - j - 1)}\right) \tag{29}$$

where the symbol $\blacktriangle$ represents one of the parameters of interest, namely, $A^*(n)$, $t_p^*(n)$, or $\tau_\alpha^*(n)$, for a time window up to k days of data. It is clear that, if the data are enough and a suitable set of parameters is found, then each of the terms in the product in (28) approaches one. Therefore, the closer the value $\gamma_k^{\bullet}$ is to one, the better the fit. Moreover, the "min" in (28) ensures that each term in the product is less than or equal to one, from which it follows that $\gamma_k^{\bullet} \leq 1$. Analyzing $\gamma_k^{\bullet}$ for different values of k enables the conclusion of whether the convergence has occurred or not within variable window sizes. We adopt windows of size 7, 14 and 21 days, in order to verify the stability of the predictions over the last one, two and three weeks.

2.6. Selecting the complexity of the model

From the previous discussion, it is possible to choose among different curve types (symmetric or asymmetric) and numbers (single or multiple sigmoids). This plays an important role both in the accuracy of the fit and in the complexity of the models (as per the different amounts of parameters to be estimated with each choice).

It should be noted that the choice of a more complex model without a significant increase in the accuracy may lead to the problem of model overfitting, that is, an exaggeration while fitting of the training data that may compromise the generalization of the model predictions [51]. In order to avoid this problem, this paper employs a generalized cross validation (GCV) approach.

The generalized cross validation index (GCVI) is defined as [52]:

$$\text{GCVI} = \text{MSE}\left(1 + \frac{2n_p}{N}\right) \tag{30}$$

where MSE is the mean squared error defined in (21), N is the number of points used to calibrate the model and n_p is the number of free parameters. Each asymmetric sigmoid contains four free parameters (A, ν, δ, t_p), whereas a symmetric sigmoid contains only three (A, δ, t_p) — note that, in a symmetric sigmoid, ν is constrained to 1.

A more complex model will generally lead to a lower MSE, but also to a higher value of n_p. The choice of the model with the minimum GCVI value allows a compromise between accuracy and complexity [52]. With this choice, a more substantial gain in the

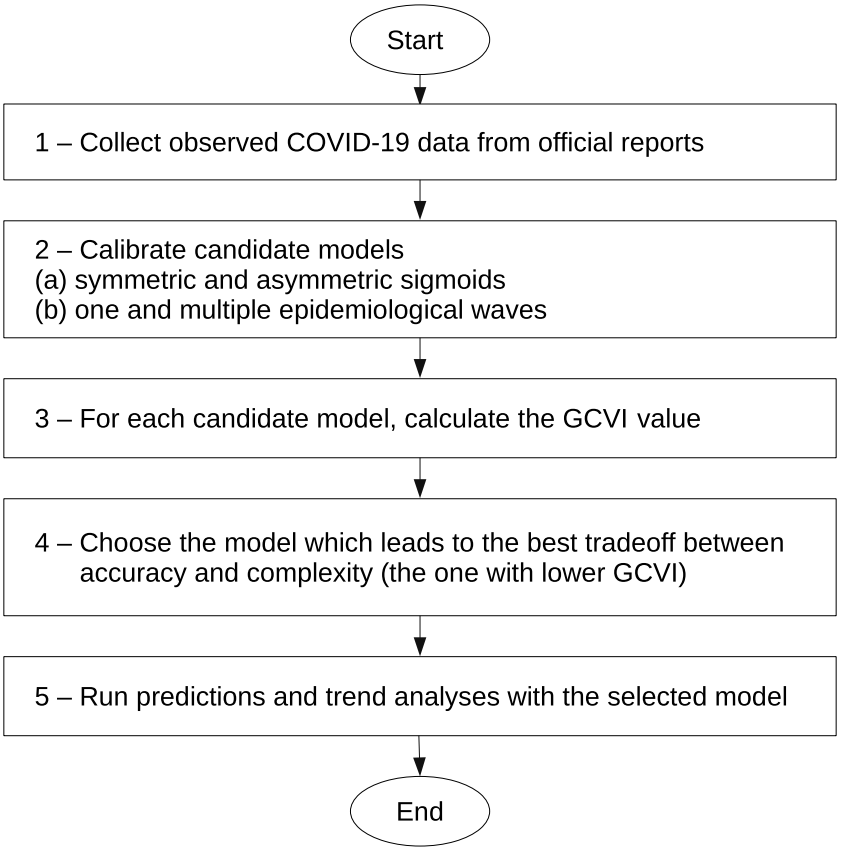


Fig. 2. Flowchart summarizing the methodology proposed in the current paper.

accuracy of the fit has to be obtained to justify a more complex curve.

Particularly for cases of regions where the contagion is in its early stage, there are not enough data to observe a deceleration phase. Therefore, in this case the data are insufficient to support estimation of the asymmetric curves. In these situations, the symmetric curves can be used in the fitting and an automated decision of whether to present results with a symmetric or an asymmetric curve is required. This decision is taken by comparing the GCVI values of the symmetric and asymmetric curves and selecting the one with the lower GCVI value.

Similarly, given the number N_s of sigmoids, two fits are performed, using either one or N_s sigmoids, and the corresponding GCVI values are calculated. The model with lower GCVI is then selected.

2.7. Uncertainty quantification

In order to quantify the uncertainty associated to the predictions, the behavior of the model in the last month is analyzed. For this purpose, the software follows the methodology explained in Section 2.6 to calculate the values of $A^*(n)$, for $n \in \{t_{end}, t_{end} - 10, t_{end} - 20, t_{end} - 30\}$, that is, the final number of occurrences predicted with data available on the last day of the analysis and on 10, 20 and 30 days prior to this date.

The ratio $\lambda_k \geq 1$ is defined as:

$$\lambda_k = \max\left(\frac{A^*(t_{end} - k)}{A^*(t_{end})}, \frac{A^*(t_{end})}{A^*(t_{end} - k)}\right) \tag{31}$$

and the overall uncertainty $\lambda \geq 1$ is described as:

$$\lambda = \max(\lambda_{30}, \lambda_{20}, \lambda_{10}) \tag{32}$$

The curves representing the minimum and maximum predicted number of occurrences are sigmoids whose parameters A_{min} and A_{max} are:

$$A_{min} = \frac{1}{\lambda} A^*(t_{end}) \tag{33}$$

$$A_{max} = \lambda\, A^*(t_{end}) \tag{34}$$

The remaining parameters (t_p, ν, δ) are the same of the nominal sigmoid.

Such sigmoids are considered only in the future, that is, for $t > t_{end}$.

2.8. Summary

The methodology proposed in the current paper is summarized by the flowchart presented in Fig. 2.

2.9. Implementation

The computer program described in this paper was developed using MATLAB® 2020a, with the Optimization Toolbox ™ and the MATLAB Compiler ™.

The program uses as data source reports published in spreadsheet format in the websites of the European Centre for Disease Prevention and Control (ECDC) [53], of Johns Hopkins University [54] and of the Brasil.io project [55].

The ECDC reports contain country-wise data of the countries in the world, while the reports of Johns Hopkins University and of the Brasil.io project presents data of United States counties and of Brazilian states and cities, respectively.

The data inform the number of newly infected and deceased people on each date. These numbers are informed separately for

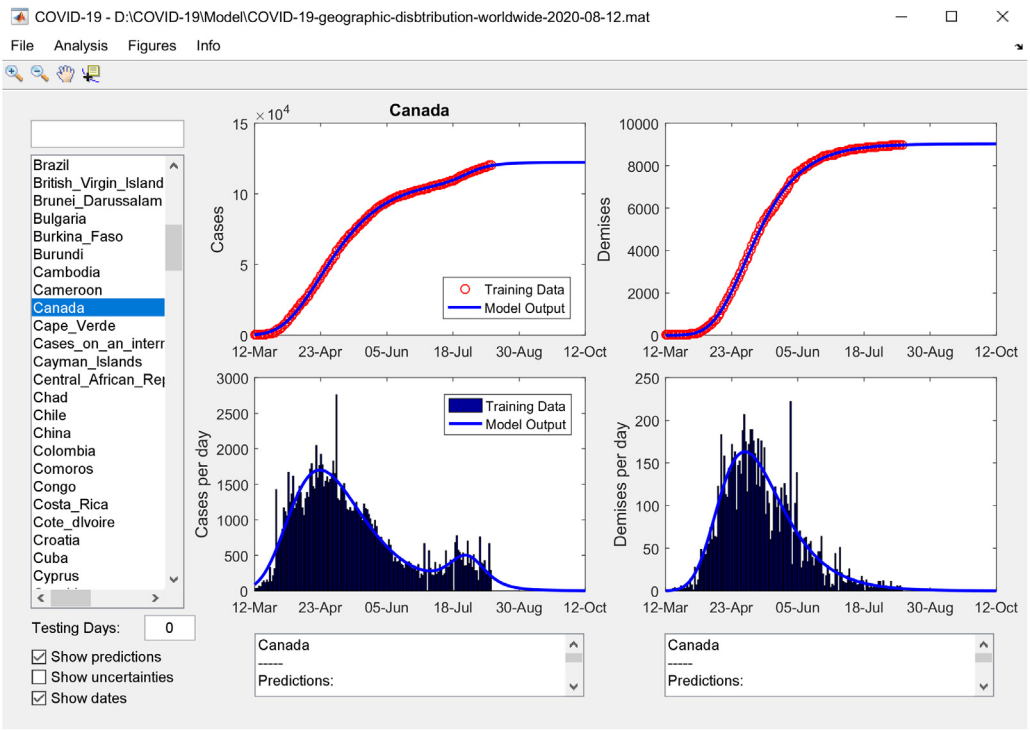


Fig. 3. Main screen of the Graphical User Interface (GUI), with worldwide data from the European Centre for Disease Prevention and Control (ECDC). Note the list of countries on the left side.

each region, allowing to perform an independent analysis for each of them.

3. Results

3.1. The graphical user interface

The computer program may be downloaded from the following link, where the data files updated until 12-Aug-2020 are also available.

https://gitlab.com/rubensjma/sigmoid-covid-19

The folder contains a "readme" file, which explains the main features and the preliminary steps to use the program. We emphasize that the program may be installed and run directly from the operating system, independently of the user possessing a licensed MATLAB® installation. Should the user have MATLAB® and the required packages installed, he/she may run directly a different file from the package without any installation.

A Graphical User Interface (GUI), illustrated in Figs. 3 and 4, will appear. These figures present the main screen of the GUI with data from the European Centre for Disease Prevention and Control and from Johns Hopkins University, respectively. A zoom was applied to these figures to allow for a better reading of their contents; that is the reason why the names of some U.S. counties appear truncated and why the predictions in the bottom of the figures appear incomplete.

When running the GUI for the first time, the user is advised to initially select the option "File: Download New Data File" of the main menu, as described in further detail below.

The GUI contains a main menu with the following options:

- **"File: Load Data File" —** This option is used to load data from a spreadsheet in the standard formats defined by the ECDC, by Johns Hopkins University and by the Brasil.io project. Upon first reading, the spreadsheet will be converted to a MATLAB® data file with extension .mat, in order to speed up the following readings. When the interface is opened, the last data file is automatically reloaded.
- **"File: Last Data Files" —** This option is used to reload one of the last data files, as illustrated in Fig. 5.
- **"File: Download New Data File"** — This option shows the menu presented in Fig. 6, where the user may select one of the following websites: ECDC, Johns Hopkins University or Brasil.io project. The user may choose to download the data file directly or to access one of these sites, using the default web browser. It is easier to choose the automatic download; however, we opt to also provide an option to access the websites as an acknowledgment of the work performed by the people responsible for them.
- **"File: Quit" —** This is a standard option to close the interface.
- **"Analysis: Run Trend Analysis" —** This option runs the trend analysis, as described in the previous section, and presents its results in an external figure. Examples of results of this analysis are presented in a following subsection.
- **"Figures: Export Figure" —** This option is used to export the graphs of the main screen to a new figure, in order to facilitate its edition and copy to external software.
- **"Figures: Close all Figures" —** This option closes all external figures.
- **"About: Info" —** This option shows updated information about the interface. It also contains an acknowledgment of the sites used as data sources.

On the left of the main screen (Figs. 3 and 4), there is a list of all available regions, which will be henceforth called the region list. In this list, when analyzing data from the ECDC or from Johns Hopkins University, the user can select the name of the desired country or of the desired U.S. county (in English); when analyzing data from the Brasil.io project, the user can select the name of the Brazilian states and cities (in Portuguese). Brazilian states are identified by their two-letter acronym. The names of the cities are presented without accents; for instance, the cities of "São

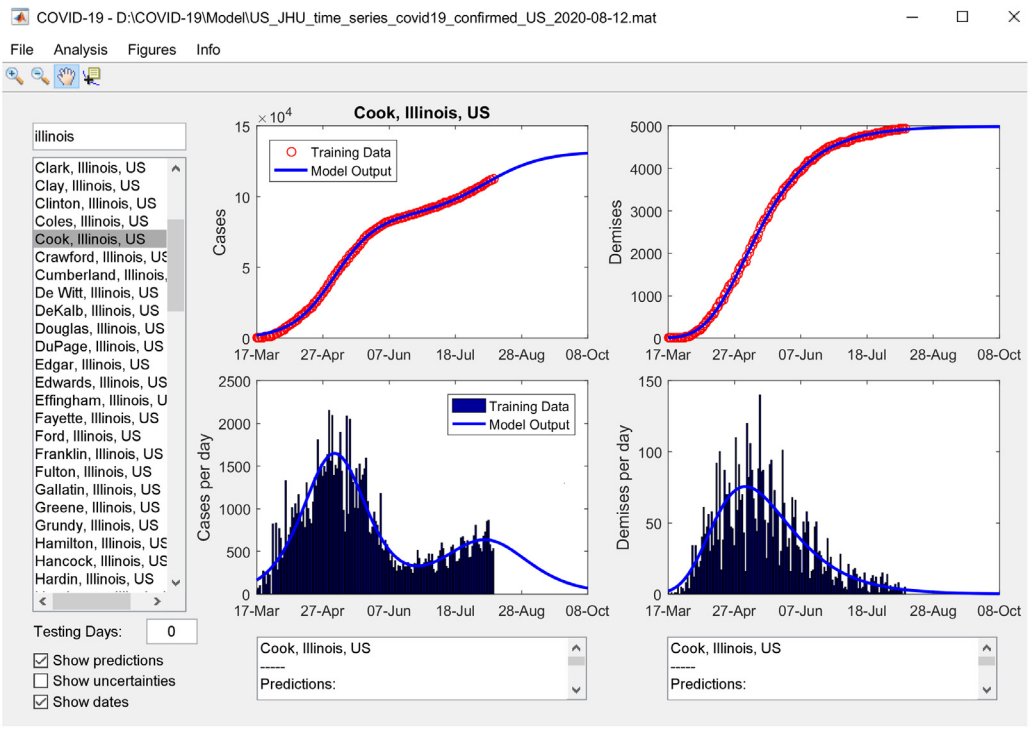


Fig. 4. Main screen of the Graphical User Interface (GUI), with U.S. data from Johns Hopkins University. Note the list of counties on the left side and the search for "Illinois" on the upper left corner.

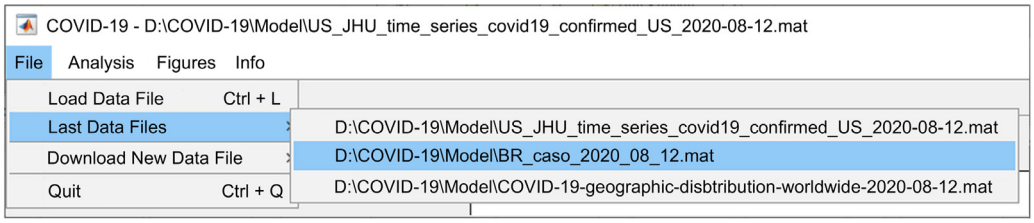


Fig. 5. "File: Last Data Files" option from the main menu.

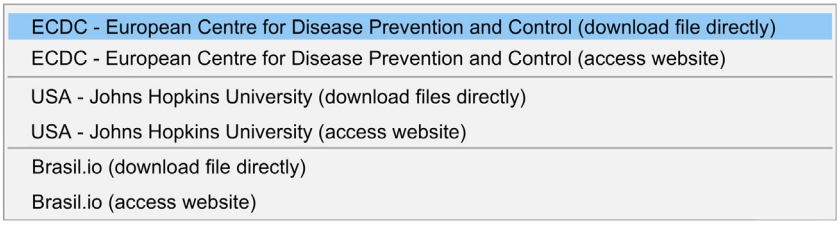


Fig. 6. Options to download new data files and to access the corresponding websites.

Paulo", "Santa Bárbara d'Oeste" and "Santa Fé" are identified as "Sao Paulo", "Santa Barbara d'Oeste" and "Santa Fe", respectively.

Above the region list, there is an edit field where the user can type the name of a region to look for on the list. The user can select the region name in full or in part, and may also employ regular expressions. Furthermore, a vertical bar "| " can be used representing the "or" operator, to perform a search for more than one region; for instance: **fran|germ|italy** will restrict the countries in the list to France, Germany and Italy. An empty string is used in the search bar to restore the complete list of regions. Note the search for "Illinois" in Fig. 4.

Below the region list, there are three options: "Show predictions", "Show uncertainties" and "Show dates". "Show predictions" is used to enable or disable the mathematical modeling (if disabled, only historical data will be shown). "Show uncertainties" is used to calculate and present the uncertainty cone. If "Show dates" is disabled, then sequential numbers are shown in the graphs' axes, instead of dates.

In the lower left corner of the GUI, there is an edit field where the user can specify the number of days for testing. For instance, if the user specifies a value of 7 days, then the model is calibrated with all data available until one week before the data acquisition, and the remaining days are used to test the model, allowing a comparison between the predictions of the model and the observed data.

The main screen presented in Figs. 3 and 4 contains four graphs, representing the accumulated (top) and daily (bottom) number of cases (left) and demises (right). Each graph contains

Table 2
Example of the information presented at the bottom of the main screen.

Canada ———— Predictions: Final number of cases: 122407 Date of maximum daily cases: 23-Apr-2020 Settling date (cases): 11-Aug-2020 ———— Number of accumulated cases on 12-Aug-2020: 120406 ———— Prediction with data from 12-Mar-2020 to 12-Aug-2020 Curve: f(t)=(109370 / (1 + 3.56e−02 exp(-((t-42.34)/23.24)))^(1/3.56e−02)) + (13037 / (1 + exp(-(t-137.34)/7.67))) Normalized RMSE (cases): 4.904e−03 R^2 (cases): 0.9998	Canada ———— Predictions: Final number of demises: 9030 Date of maximum daily demises: 02-May-2020 Settling date (demises): 18-July-2020 ———— Number of accumulated demises on 12-Aug-2020: 8991 ———— Prediction with data from 12-Mar-2020 to 12-Aug-2020 Curve: f(t)=(9030 / (1 + 5.93e−02 exp(-((t-51.30)/19.77)))^(1/5.93e−02)) Normalized RMSE (demises): 5.499e−03 R^2 (demises): 0.9998

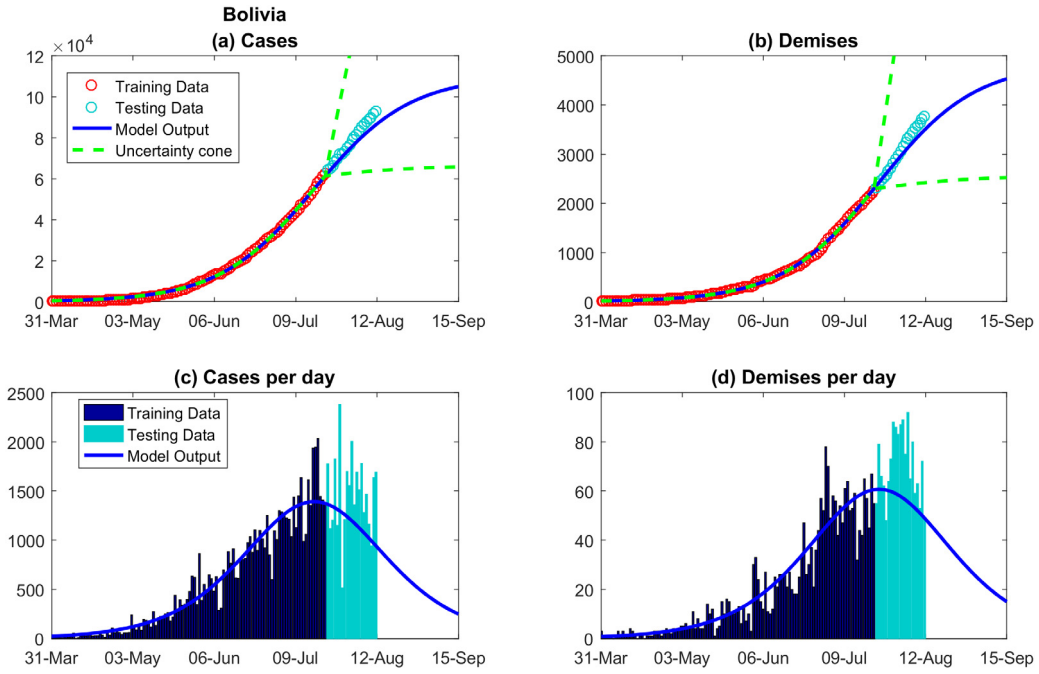


Fig. 7. Graphs of cases and demises for the country of Bolivia. The figure shows accumulated (top) and daily (bottom) occurrences.

historical data and theoretical curves representing them. Observed data are presented using either circles (accumulated values) or bars (daily values). The model output is represented by the continuous blue line.

Below the graphs, the following predictions are presented, for either cases or demises: final number, date of maximum daily occurrences and settling date (as defined in Table 1). Furthermore, the equation of the best sigmoid (or set of sigmoids in case more than one wave is identified) matching the accumulated data is presented, as well as the indices RMSE and R^2, defined in the previous section. The predictions are presented in editable fields, so that the user can copy their texts and paste them in external software. An example of such information is presented in Table 2.

3.2. Illustrative results — time series

In order to illustrate the use of the tool to perform predictions, Figs. 7 to 9 show the model results for the country of Bolivia, the U.S. county of Los Angeles and the Brazilian capital city of Brasilia, respectively. Data updated on 12-Aug-2020 were used. The number of testing days was set to 21, meaning that the model was calibrated with data until 21-July-2020 and the following three weeks were used to compare the predicted and observed values.

3.3. Illustrative results — trend analysis

As previously mentioned, the trend analysis is run when the user selects the corresponding option in the main menu. Examples of figures resulting from such analysis are presented in Figs. 10 to 12, which correspond to the Brazilian city of São Paulo (SP), the US county of Cook, Illinois, and the Brazilian state of São Paulo, respectively. Each of these figures contains three subfigures, showing the predicted value of the three parameters of interest described in Table 1.

The abscissa of the graphs indicates the date of the estimation, meaning that all data available until that date were used to estimate the value of the parameter under study. It can be seen that, as expected, the values of the estimated parameters vary with the amount of data used to estimate them.

On the title of each subfigure, the values of γ_7, γ_{14} and γ_{21} are presented, indicating how stable each prediction is, considering the last one, two and three weeks, respectively. A value of γ_k closer to one indicates a more stable prediction.

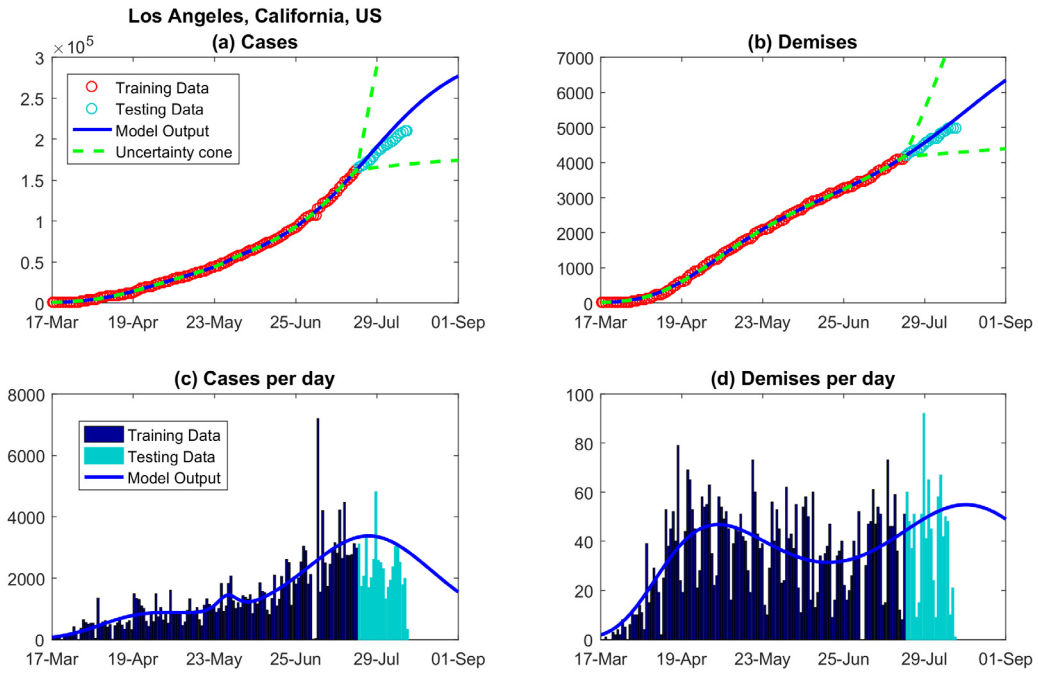


Fig. 8. Graphs of cases and demises for the U.S. county of Los Angeles. The figure shows accumulated (top) and daily (bottom) occurrences.

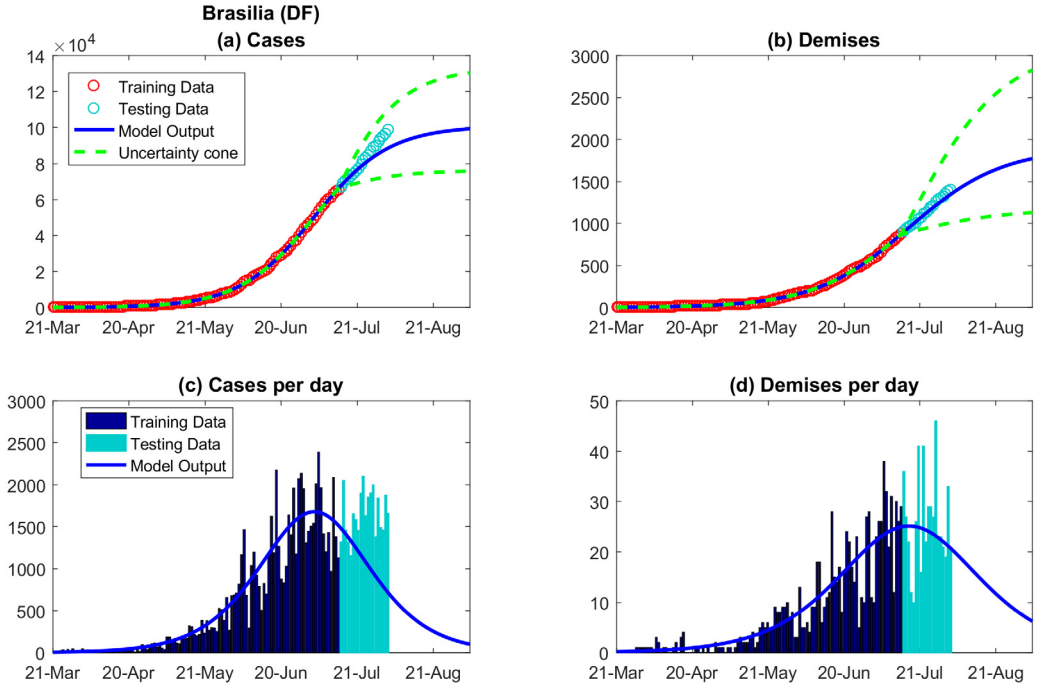


Fig. 9. Graphs of cases and demises for the Brazilian capital city of Brasilia. The figure shows accumulated (top) and daily (bottom) occurrences.

4. Discussion

Figs. 7 to 9 indicate that the model represents well the training data and that the observed accumulated values (subfigures (a) and (b)) follow closely the values predicted by the model. In the first week ahead, the observed results are very close to the predicted ones. In the following weeks, the results are still close, but the observed values start to drift away from the model outputs, although inside the established tolerance. In fact, forecasts are expected to diminish in accuracy over time. Nevertheless, as will be discussed below, our methodology presents parameters to assess the quality of the predictions and to analyze their trend.

When analyzing the daily experimental curves (subfigures (c) and (d)), it can be seen that there are high amplitude fluctuations, which may be ascribed to the nature of the observed data and may be associated to non-uniform delays in the official notifications of contaminations and deceases. These fluctuations in amplitude are similar to the sensor noise observed when analyzing physical data. In the study of dynamical systems, it is known that integral operations are robust to the presence of noise. By analogy, since our methodology uses the accumulated data (and

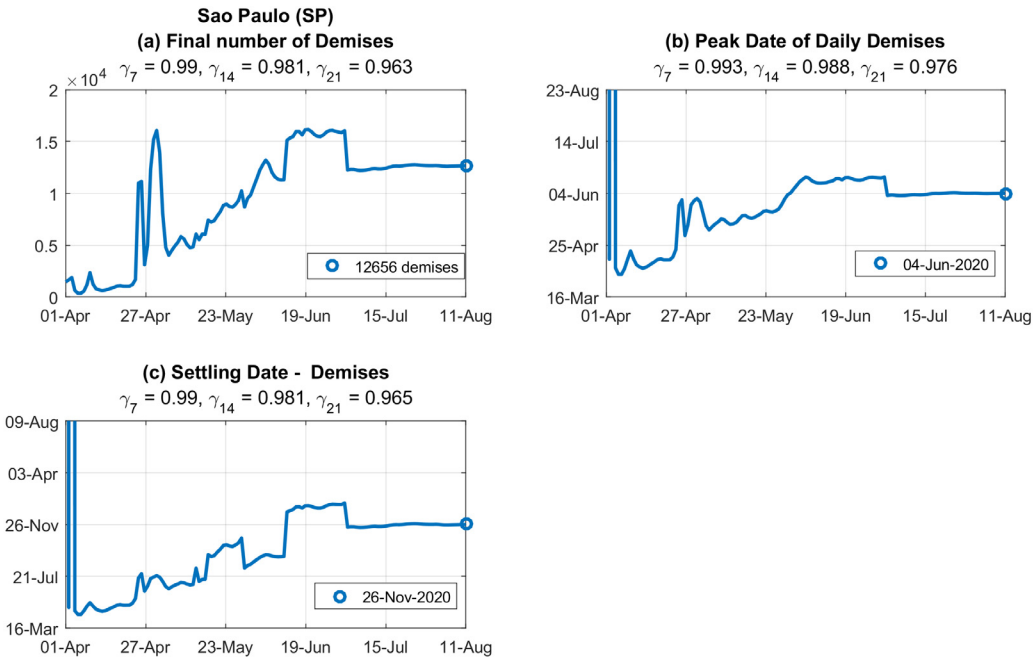


Fig. 10. Trend analysis results for the Brazilian city of São Paulo (SP). The abscissa indicates the date of the estimation.

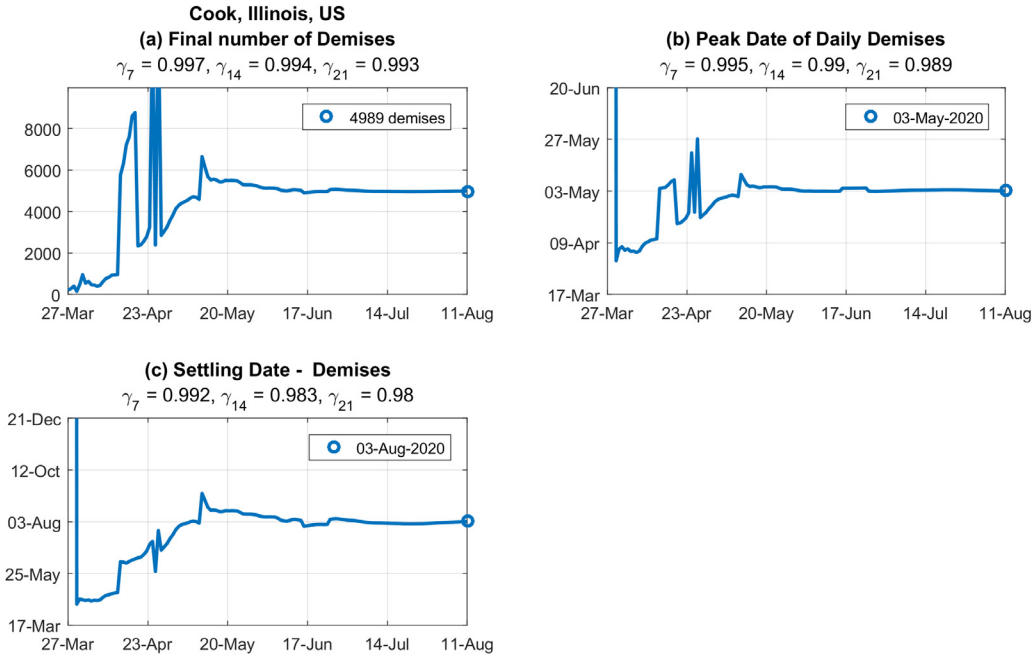


Fig. 11. Trend analysis results for the U.S. county of Cook, Illinois. The abscissa indicates the date of the estimation.

not the daily values) to estimate the model parameters, it can be concluded that it is less sensitive to daily fluctuations in the data.

The proposed model is able to identify and represent as many epidemiological waves as necessary. For instance, the two peaks observed in the model representation of the number of daily cases in Fig. 4 indicate a clear occurrence of two epidemiological waves in the U.S. county of Cook, Illinois — the daily cases were decreasing until the second week of June, and then started to consistently increase again. On the other hand, only one wave is observed in the Brazilian city of Brasilia (Fig. 9(c)). To the best of the authors' knowledge, no more than two epidemiological waves have been reported in any region yet. However, it may still happen, especially if there are frequent changes in the public policies, imposing and relieving containment actions. The model is ready to represent this behavior.

Figs. 10 to 12, with the results of the trend analysis, indicate how the prediction of each parameter of interest varies with time.

Fig. 10(a) represents the final number of predicted demises in the city of São Paulo (SP). It can be seen that there are oscillations in the predictions until May 5th. These oscillations result from the inclusion of new data and are expected to occur when the pandemic is at an early stage. From May 5th to June 12th, there is a clear increasing tendency in the number of demises. On June 13th, the prediction stabilizes around approximately 15000 demises. Finally, on July 2nd, the number of demises reaches approximately 12600 and is stabilized around this value ever

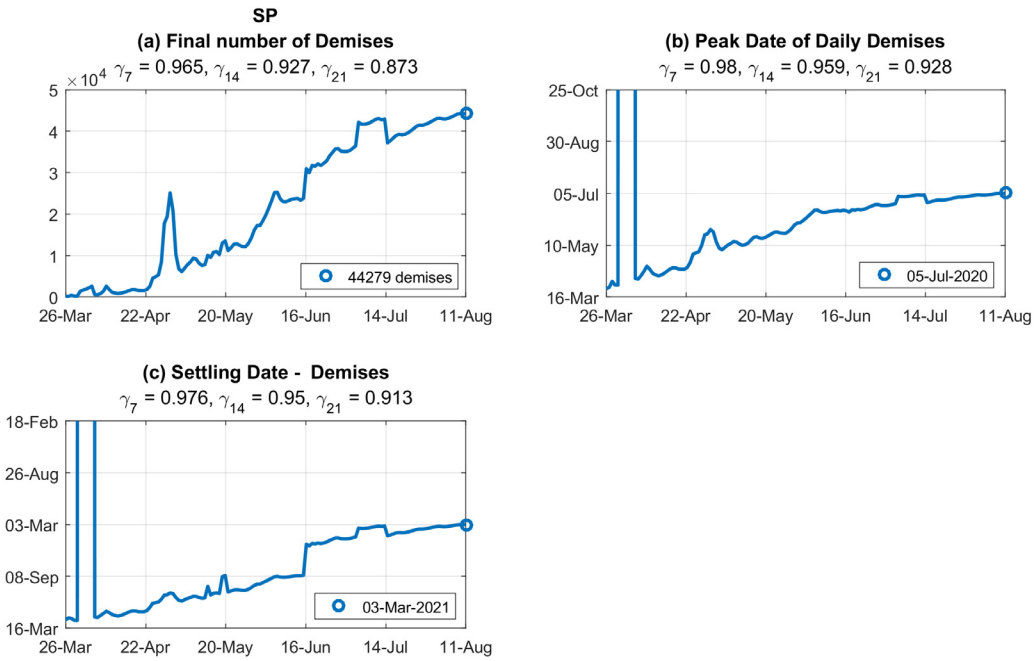


Fig. 12. Trend analysis results for the Brazilian state of SP (São Paulo). The abscissa indicates the date of the estimation.

since. Likewise, Figs. 10-(b) and (c) indicate a stable prediction of the date of maximum daily demises and of the settling date since July 2nd.

Similarly, Fig. 11-(a) indicates that the dynamics of the pandemic in the U.S. county of Cook, Illinois, followed a similar pattern. There were oscillations until the end of April, followed by an increasing trend until May 17th, a slightly decreasing tendency and finally an stabilized predicted value of approximately 5000 demises since June 17th.

On the other hand, Fig. 12 indicates that the pandemic is not yet stabilized in the Brazilian state of SP (São Paulo). An increasing tendency can be seen over the last weeks. For instance, Fig. 12-(a) shows that the predicted number of demises was approximately 37000 on July 14th and changed to 44000 on August 11th, indicating an increase of approximately 20% in 28 days.

The stability of such predictions over the last one, two and three weeks may be verified by the values of γ_7, γ_{14} and γ_{21} shown in the title of each figure. It can be seen that these values are close to one, indicating stabilized or near-to-stabilization predictions.

The values of γ_k are intended to represent the convergence of the estimations. As an additional feature, they may also be used as a measurement of the quality of the prediction — higher values of γ indicate that the pandemic has been following the same predicted behavior over the last weeks. For instance, when analyzing the values of γ_{21} presented in Figs. 10-(a) and 12-(a), it can be seen that such values are $\gamma_{21} = 0.963$ and $\gamma_{21} = 0.873$ for the city and for the state of São Paulo, respectively. One may infer from these numbers that the predictions obtained in the last three weeks are more stable in the city of São Paulo than in the state with the same name. This is the same conclusion that was achieved by analyzing the curves, as described in the previous paragraphs.

It should be emphasized that the values in Figs. 10-(a), 11-(a) and 12-(a) do not refer to the number of demises on the day of the analysis, but rather to the predicted final number of demises, estimated on the basis of all data available until that date. This is an innovative approach for trend analysis in this context and, to the best of the authors' knowledge, has not been proposed before.

Additionally, the same analyses presented here for the number of demises can be run for the number of infected people.

Typical trends observed in this kind of analysis are (a) oscillations, (b) increasing values and (c) stabilized values. High-amplitude oscillations may occur in the beginning of the pandemic and do not allow to reach any conclusion; however, they usually disappear after the first few weeks. Increasing values indicate a need of more compelling action by the authorities, while stabilized values indicate that the pandemic is under control.

The stabilized results for the city of São Paulo and for the county of Cook, Illinois, allow to conclude that the actions of the local governments to control the pandemic are taking effect. It is of public interest to determine how the disease will spread in each city after the restriction measures are alleviated. For this purpose, the trend analysis should be run again. Should a new increasing tendency be observed, the authorities would be advised to reinstate some containment measures.

It is important to point out that these results, although helpful, should be validated by medical experts and not be considered alone when deciding public policies.

The trend analysis may be run for a country, a state, a county or a city. It provides more useful information when it is run for smaller administrative regions such as a county or a city, because it allows supporting decision by local authorities based on specific data of the region under consideration.

A limitation of the proposed approach is that it is not adequate to analyze the pandemic in very small cities or counties, because the number of infections and demises is usually very low, not allowing a good fitting by the mathematical model proposed here. However, for medium- and large-sized cities or counties, informative results are expected, as the ones presented here for the U.S. counties of Cook, Illinois and Los Angeles, California and for the Brazilian cities of Brasilia (DF) and São Paulo (SP).

Moreover, caution must be taken in using the forecast capability, especially for longer time intervals. For instance, the occurrence of a second wave can be detected with our proposed method only once enough data is available, i.e., forecasting a

second wave is not possible. After detection, enough data must be fed to perform the fitting to achieve a trustworthy estimation of the magnitude of this eventual new wave, similarly to the analysis in the beginning of a first wave that we showed. Therefore, we do not advise the usage of the tool to support claims that the pandemic is over. It must remain as an auxiliary tool to assess the number of infected/deceased individuals in a short period after the last fit, as shown in our results. We also emphasize that updating the model periodically is recommended.

The results presented here are illustrative and correspond to the scenario on the date when the data were acquired, that is, on 12-Aug-2020. These analyses should always employ updated data to increase their reliability. Therefore, the authors recommend these studies to be repeated periodically, at least on a weekly basis. The developed computer program allows to easily perform this task.

5. Conclusion

This paper proposed a methodology and a computational tool to forecast the COVID-19 pandemic throughout the world, providing useful resources for health-care authorities. A user-friendly Graphical User Interface (GUI) in MATLAB® was developed and can be downloaded online for free use. An innovative approach for trend analysis was presented.

Resources in the computational tool allow to quickly run analyses for the desired regions. Additional options allow to access the official website of the European Centre of Disease Prevention and Control, of Johns Hopkins University and of the Brasil.io project, in order to download new data as soon as they are published online. To this date, these institutions have been updating their reports on a daily basis.

The analyses run by the program are intended only as an aid and the results should be interpreted with care. They do not replace a careful analysis by experts. Nevertheless, such results may be a very useful tool to assist the authorities in their decision-making process.

The proposed program is in continuous development and future added features will be published and described in the project webpage. The authors would appreciate any feedback and suggestions to improve the computational tool.

The program, in its current version, is able to process detailed information about U.S. counties and about Brazilian states and cities. These two countries were chosen because they have continental dimensions and are currently the focus of the COVID-19 pandemic. Nevertheless, the same resource could be extended to other countries. For this purpose, the main requirement would be to write a code to read other country data files and convert them to the format recognized by the program, which is quite simple.

Future works can employ the same methodology and adapt the computer tool to describe the dynamics of other epidemics around the world. In the recent past, no pandemic was as severe as the COVID-19, but there were occurrences of other diseases such as Influenza A and MERS-CoV. Should a similar epidemic occur again, the computer program described here would be a resourceful tool.

CRediT authorship contribution statement

Mohallem Paiva: Conceptualization, Data curation, Methodology, Project administration, Software, Validation, Formal analysis, Investigation, Writing, Visualization, Supervision.
Rubens Junqueira Magalhães Afonso: Conceptualization, Data curation, Methodology, Validation, Formal analysis, Investigation, Writing, Visualization.
Fabiana Mara Scarpelli de Lima Alvarenga Caldeira: Formal analysis, Writing.
Ester de Andrade Velasquez: Formal analysis, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the European Centre for Disease Prevention and Control (ECDC), Johns Hopkins University and the Brasil.io project for making the COVID-19 data publicly available and for allowing its use for research purposes.

Rubens Afonso acknowledges the support of CAPES, Brazil (fellowship proc. #88881.145490/2017-01) and the Federal Ministry for Education and Research of Germany through the Alexander von Humboldt Foundation, Germany.

Henrique Paiva acknowledges the support of the Sao Paulo Research Foundation FAPESP, Brazil.

References

[1] C. Huang, Y. Wang, X. Li, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (10223) (2019) 497–506, http://dx.doi.org/10.1016/S0140-6736(20)30183-5 (published correction appears in Lancet. 30 January 2020).
[2] World Health Organization (WHO), Novel coronavirus – China, 2020, http://www.who.int/csr/don/12-January-2020-novel-coronavirus-china/en/. (Accessed 09 May 2020).
[3] H.Y. Geng, W.J. Tan, A novel human coronavirus: Middle East respiratory syndrome human coronavirus, Sci. China Life Sci. 56 (8) (2013) 683–687, http://dx.doi.org/10.1007/s11427-013-4519-8.
[4] N.M. Linton, T. Kobayashi, Y. Yang, et al., Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data, J. Clin. Med. 9 (2) (2020) 538, http://dx.doi.org/10.3390/jcm9020538.
[5] World Health Organization (WHO), Coronavirus disease (covid-19). Situation report – 51, 2020, http://www.who.int/docs/default-source/coronaviruse/situation-reports. (Accessed 09 May 2020).
[6] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, Lancet 395 (10225) (2019) 689–697, http://dx.doi.org/10.1097/01.ogx.0000688032.41075.a8 (published correction appears in Lancet. 4 February 2020).
[7] M. Chinazzi, J.T. Davis, M. Ajelli, et al., The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak, Science 368 (6489) (2019) 395–400, http://dx.doi.org/10.1126/science.aba9757.
[8] World Health Organization (WHO), COVID-19 weekly epidemiological update - 24 November 2020, 2020, https://www.who.int/publications/m/item/weekly-epidemiological-update---24-november-2020. (Accessed 30 November 2020).
[9] L. Chen, J. Xiong, L. Bao, Y. Shi, Convalescent plasma as a potential therapy for COVID-19, Lancet Infect Dis. 20 (4) (2020) 398–400, http://dx.doi.org/10.1016/s1473-3099(20)30141-9.
[10] M. Guastalegname, A. Vallone, Could chloroquine /hydroxychloroquine be harmful in coronavirus disease 2019 (COVID-19) treatment? Clin. Infect. Dis. (2019) 321, http://dx.doi.org/10.1093/cid/ciaa321 (published online ahead of print, 24 March 2020).
[11] K. Lundstrom, Coronavirus pandemic - therapy and vaccines, Biomedicines 8 (2020) 109, http://dx.doi.org/10.3390/biomedicines8050109.
[12] M. Sarkar, A.S. Agrawal, R.S. Dey, S. Chattopadhyay, R. Mullick, P. De, S. Chakrabarti, M. Chawla-Sarkar, Molecular characterization and comparative analysis of pandemic H1N1/2009 strains with co-circulating seasonal H1N1/2009 strains from eastern India, Arch. Virol. 156 (2) (2009) 207–217, http://dx.doi.org/10.1007/s00705-010-0842-6.
[13] R. Ross, CDC Estimate of Global H1N1 Pandemic Deaths: 284,000, Center for Infectious Disease Research and Policy, 2012, https://www.cidrap.umn.edu/news-perspective/2012/06/cdc-estimate-global-h1n1-pandemic-deaths-284000. (Accessed 11 May 2020).
[14] Y. Kim, S. Lee, C. Chu, S. Choe, S. Hong, Y. Shin, The characteristics of Middle Eastern respiratory syndrome coronavirus transmission dynamics in South Korea, Osong Public Health Res. Perspect. 7 (1) (2016) 49–55, http://dx.doi.org/10.1016/j.phrp.2016.01.001.

[15] J.F. Chan, S. Sridhar, C.C. Yip, S.K. Lau, P.C. Woo, The role of laboratory diagnostics in emerging viral infections: the example of the Middle East respiratory syndrome epidemic, J. Microbiol. 55 (3) (2017) 172–182, http://dx.doi.org/10.1007/s12275-017-7026-y.
[16] A.F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, R.E. Rothman, Influenza forecasting with Google flu trends, PLoS One (2013) 8, http://dx.doi.org/10.1371/journal.pone.0056176.
[17] H. Nishiura, Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (H1N1-2009), BioMed EngOnLine 10 (2009) 15, http://dx.doi.org/10.1186/1475-925x-10-15.
[18] J.P. Chretien, D. George, J. Shaman, R.A. Chitale, F.E. McKenzie, Influenza forecasting in human populations: a scoping review, PLoS One (2014) 9, http://dx.doi.org/10.1371/journal.pone.0094130.
[19] A. Longobardi, P. Villani, Trend analysis of annual and seasonal rainfall time series in the Mediterranean area, Int. J. Climatol. 30 (10) (2010) 1538–1546, http://dx.doi.org/10.1002/joc.2001.
[20] R. Atan, S.A. Raman, M.S. Sawiran, N. Mohamed, R. Mail, Financial performance of Malaysian local authorities: A trend analysis, in: 2010 International Conference on Science and Social Research, 2010, pp. 271–276, http://dx.doi.org/10.1109/cssr.2010.5773782.
[21] M. Wen, P. Li, L. Zhang, Y. Chen, Stock market trend prediction using high-order information of time series, IEEE Access 7 (2019) 28299–28308, http://dx.doi.org/10.1109/access.2019.2901842.
[22] P.T. Oliveira, C.M. Santos e Silva, K.C. Lima, Climatology and trend analysis of extreme precipitation in subregions of Northeast Brazil, Theor. Appl. Climatol. 130 (1–2) (2017) 77–90, http://dx.doi.org/10.1007/s00704-016-1865-z.
[23] J. Zhao, T. Zuo, R. Zheng, S. Zhang, H. Zeng, C. Xia, W. Chen, Epidemiology and trend analysis on malignant mesothelioma in China, Chin. J. Cancer Res. 29 (4) (2017) 361, http://dx.doi.org/10.21147/j.issn.1000-9604.2017.04.09.
[24] S.C.M. Soares, K.M.R. dos Santos, F.C.G. de Morais Fernandes, I.R. Barbosa, D.L.B. de Souza, Testicular cancer mortality in Brazil: trends and predictions until 2030, BMC Urol. 19 (1) (2019) 59, http://dx.doi.org/10.1186/s12894-019-0487-z.
[25] B. Zahmatkesh, A. Keramat, N. Alavi, A. Khosravi, A. Kousha, A.G. Motlagh, R. Chaman, Breast cancer trend in Iran from 2000 to 2009 and prediction till 2020 using a trend analysis method, Asian Pac. J. Cancer Prev. 17 (3) (2000) 1493–1498, http://dx.doi.org/10.7314/apjcp.2016.17.3.1493.
[26] A. Mousavizadeh, M. Dastoorpoor, E. Naimi, K. Dohrabpour, Time-trend analysis and developing a forecasting model for the prevalence of multiple sclerosis in Kohgiluyeh and Boyer-Ahmad Province, southwest of Iran, Public Health 154 (2018) 14–23.
[27] H. Yuan, X. Li, G. Wan, L. Sun, X. Zhu, F. Che, Z. Yang, Type 2 diabetes epidemic in East Asia: a 35-year systematic trend analysis, Oncotarget 9 (6) (2018) 6718, http://dx.doi.org/10.18632/oncotarget.22961.
[28] Q. Lin, S. Zhao, D. Gao, Y. Lou, S. Yang, S.S. Musa, M. Wang, Y. Cai, W. Wang, L. Yang, D. He, A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action, Int. J. Infect. Dis. 93 (2019) 211–216, http://dx.doi.org/10.1016/j.ijid.2020.02.058.
[29] H.M. Paiva, R.J.M. Afonso, I.L. de Oliveira, G.F. Garcia, A data-driven model to describe and forecast the dynamics of COVID-19 transmission, PLoS One 15 (7) (2020) e0236386, http://dx.doi.org/10.1371/journal.pone.0236386.
[30] C. Hou, J. Chen, Y. Zhou, L. Hua, J. Yuan, S. He, J. Zhang, The effectiveness of quarantine of wuhan city against the Corona virus disease 2019 (COVID-19): A well-mixed SEIR model analysis, J. Med. Virol. (2020) http://dx.doi.org/10.1002/jmv.25827.
[31] R. Salgotra, M. Gandomi, A.H. Gandomi, Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming, Chaos Solitons Fractals (2020) 109945, http://dx.doi.org/10.1016/j.chaos.2020.109945.
[32] M.H.D.M. Ribeiro, R.G. da Silva, V.C. Mariani, L. dos Santos Coelho, Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil, Chaos Solitons Fractals (2020) 109853, http://dx.doi.org/10.1016/j.chaos.2020.109853.
[33] Zifeng Yang, et al., Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, J. Thorac. Dis. 12 (3) (2020) 165, http://dx.doi.org/10.21037/jtd.2020.02.64.
[34] A. Hernandez-Matamoros, H. Fujita, T. Hayashi, H. Perez-Meana, Forecasting of COVID19 per regions using ARIMA models and polynomial functions, Appl. Soft Comput. (2020) 106610, http://dx.doi.org/10.1016/j.asoc.2020.106610.
[35] S. Rath, A. Tripathy, A.R. Tripathy, Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model, Diabetes Metab. Syndr. Clin. Res. Rev. (2020) http://dx.doi.org/10.1016/j.dsx.2020.07.045.
[36] Y.F. Lin, Q. Duan, Y. Zhou, T. Yuan, P. Li, T. Fitzpatrick, et al., Spread and impact of COVID-19 in China: a systematic review and synthesis of predictions from transmission-dynamic models, Front. Med. 7 (2020) 321, http://dx.doi.org/10.3389/fmed.2020.00321.
[37] Y. Mohamadou, A. Halidou, P.T. Kapen, A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19, Appl. Intell. (2020) 1–13, http://dx.doi.org/10.1007/s10489-020-01770-9.
[38] F.J. Richards, A flexible growth function for empirical use, J. Exp. Bot. 10 (2) (1959) 290–301, http://dx.doi.org/10.1093/jxb/10.2.290.
[39] R.C. Dorf, R.H. Bishop, Modern Control Systems, thirteenth ed., Pearson, London, 2016.
[40] P.E. Gill, E. Wong, Sequential quadratic programming methods, in: J. Lee, S. Leyffer (Eds.), Mixed Integer Nonlinear Programming, Springer, New York, 2012, pp. 147–224, http://dx.doi.org/10.1007/978-1-4614-1927-3_6.
[41] W.U. Khan, Z. Ye, N.I. Chaudhary, M.A.Z. Raja, Backtracking search integrated with sequential quadratic programming for nonlinear active noise control systems, Appl. Soft Comput. 73 (2018) 666–683, http://dx.doi.org/10.1016/j.asoc.2018.08.027.
[42] S. Khalilpourazari, S.H.R. Pasandideh, S.T.A. Niaki, Optimization of multi-product economic production quantity model with partial backordering and physical constraints: SQP, SFS, SA, and WCA, Appl. Soft Comput. 49 (2016) 770–791, http://dx.doi.org/10.1016/j.asoc.2016.08.054.
[43] J. Nocedal, S.J. Wright, Numerical Optimization, second ed., Springer, New York, 2006, http://dx.doi.org/10.1007/b98874.
[44] Mathworks, Documentation of the fmincon function, 2020, Available at https://www.mathworks.com/help/optim/ug/fmincon.html. (Accessed 12 August 2020).
[45] A Aleta, D. Martín-Corral, A. Pastore y Piontti, et al., Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19, Nat. Hum. Behav. (2020) http://dx.doi.org/10.1038/s41562-020-0931-9.
[46] K. Leung, J.T. Wu, D. Liu, G.M. Leung, First-wave COVID-19 transmissibility and severity in China outside hubei after control measures, and second-wave scenario planning: a modelling impact assessment, Lancet 395 (10223) (2020) 1382–1393, http://dx.doi.org/10.1016/s0140-6736(20)30746-7.
[47] L. López, X. Rodó, The end of social confinement and COVID-19 re-emergence risk, Nat. Hum. Behav. 4 (2020) 746–755, http://dx.doi.org/10.1038/s41562-020-0908-8.
[48] S. Xu, Y. Li, Beware of the second wave of COVID-19, Lancet 395 (10233) (2020) 1321–1322, http://dx.doi.org/10.1016/S0140-6736(20)30845-X.
[49] J.C. Lagarias, J.A. Reeds, M.H. Wright, P.E. Wright, Convergence properties of the Nelder–Mead simplex method in low dimensions, SIAM J. Optim. 9 (1) (1999) 112–147, http://dx.doi.org/10.1137/S1052623496303470.
[50] A.G. Barnston, Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score, Weather Forecast. (1992) 699–709, http://dx.doi.org/10.1175/1520-0434(1992)007<0699:CATCRA>2.0.CO;2.
[51] E.W. Steyerberg, Overfitting and optimism in prediction models, in: Clinical Prediction Models, Springer, Cham, 2019, pp. 95–112, http://dx.doi.org/10.1007/978-0-387-77244-8_5.
[52] H.M. Paiva, R.K.H. Galvão, Wavelet-packet identification of dynamic systems in frequency subbands, Signal Process. 86 (8) (2006) 2001–2008, http://dx.doi.org/10.1016/j.sigpro.2005.09.021.
[53] European Centre for Disease Prevention and Control (ECDC), Download today's data on the geographic distribution of COVID-19 cases worldwide, 2020, https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. (Accessed 12 August 2020).
[54] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. 20 (5) (2020) 533–534, http://dx.doi.org/10.1016/s1473-3099(20)30120-1.
[55] Brasil.io Project, 2020, http://brasil.io/. (Accessed 12 August 2020).