Self-consistent Estimation of Ordinary Differential Equation Parameters Describing Dynamical Systems: A Case Study of COVID-19 in Germany

LOUKAS KYRIAKIDIS[1], MICHAIL KYRIAKIDIS[2]

[1]Institute of Low-Carbon Industrial Processes
German Aerospace Center, Walther-Pauer-Straße 5, Cottbus 03046,
GERMANY

[2]Department of Technology of Digital Industry/Department of Informatics & Telecommunications
University of Athens/National Center for Scientific Research "DEMOKRITOS"
Psachna 34400/Gregoriou E & 27 Neapoleos Str., Athens 15341
GREECE

Abstract: Nowadays, the estimation of parameters for ordinary differential equations (ODEs) from historical data (time series) in optimization problems presents various challenges. These challenges include convergence to local minima when applying traditional optimization methods, inaccurate integration methods of ODEs during the optimization process, and inaccurate cost functions. To address these issues, we propose a novel methodology for estimating the parameters of ODEs that describe dynamic systems in fields such as biological populations, disease spread (e.g., COVID-19). Our methodology is based on the integration of trajectory simulation, optimization of a cost function using noisy data, and heuristic search algorithms such as genetic algorithms for minimization. We demonstrate the effectiveness of this methodology through one use case in this work: the evolution of the COVID-19 disease in German society during the first wave. The results show a highly accurate methodology capable of reproducing real-world curves with high precision.

Key-Words: Parameter Estimation of ODEs, Genetic Algorithm, FIML, COVID-19

Received: March 14, 2024. Revised: August 13, 2024. Accepted: September 15, 2024. Available online: October 22, 2024.

1 Introduction

Time series analysts assert the potential for extracting more information from dynamical processes. By observing various variables related to dynamical systems, it becomes possible to find out the governing laws dictating the time evolution of these variables. This allows for the deciphering and acquisition of knowledge about the dynamical systems and processes themselves. ODEs play a crucial role in modeling dynamical processes across various fields such as science, engineering, and medicine, not only in academia but also increasingly in industry and commerce.

Establishing such models involves the integration of theory with experimental or observational data, which includes the task of determining model parameters to best replicate the data. Quantitatively correct models are particularly important when these models are further employed in design optimization, optimal control, or forecasting [1].

Methods based on global minimization routines including random search and adaptive stochastic methods [2, 3, 4, 5], clustering methods [6], evolutionary computation [7], simulated annealing, and heuristic search such as genetic algorithms [8, 9, 10, 11] provide a detailed discussion of these methods regarding parameter identification in ODEs, while [12] and [13] offer comprehensive discussions on genetic algorithms. The disadvantage of stochastic optimizers is mainly their immense computational cost, which is the price for the flexibility and stability of these methods.

Additionally, local optimization procedures such as Newton and quasi-Newton methods [14] are computationally efficient. However, they tend to converge to local minima or become numerically unstable due to multicollinearity, particularly when the independent variables are highly correlated with each other, and the matrices used for parameter computation are near singular.

In the case of parameter identification in ODEs, one approach involves leveraging the fact that the trajectory is uniquely determined by the parameters and initial values. This can be achieved by maximizing a maximum-likelihood functional or minimizing a cost function (utility, cost or objective function).

Compared to the initial value approach, multiple

shooting offers enhanced stability with only a slight increase in computational cost, particularly for complex functions with more maxima and/or minima. The method was originally introduced by [15] and was subsequently enhanced and mathematically analyzed by [1, 16, 17, 18, 19]. Recent works [20, 21, 22, 23] have also explored the use of neural networks and certainty-equivalent expectation maximization to forecast the spread of COVID-19. The cost function utilized in the aforementioned papers, as well as by [24], is either the sum or a weighted sum of squared errors. The method, commonly referred to as regression, relies on the crucial assumption that the independent residuals fulfill the multivariate Gaussian distribution when regressing noisy variables. However, this assumption does not hold true when dealing with time series variables, necessitating the implementation of certain corrections.

In this paper, we employ the determinant of the variance-covariance matrix as the cost function to be minimized. This determinant is computed from the opposite of the "Full Information Maximum Likelihood" (FIML) functional, which ensures consistency between the residuals derived from the estimated derivatives—based on the single shooting initial value method for integrating ODEs and parameters obtained via a genetic algorithm—and the actual observed values. In the case of autocorrelated residuals, we adjust for this by incorporating autoregressive terms. With this approach and with the use of a genetic algorithm for minimization, we estimate more realistic values for the ODEs' parameters.

In the following sections, we present our methodology and examine the spread of COVID-19 disease in societies (specifically focusing on the first wave of COVID-19 in Germany) by applying the Bass diffusion model and the SIR model. Finally, this paper ends with a conclusion and provides an outlook for future work.

2 Methodology

In this work, we consider dynamic systems described by Ordinary Differential Equations (ODEs), which can be written as follows:

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} - \boldsymbol{f}(\boldsymbol{x}(t), t, \boldsymbol{\theta}, \boldsymbol{z}(t)) = 0, \;\; \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \quad (1)$$

where $\boldsymbol{x}(t) \in \mathbb{R}^d$ denotes the d-dimensional state vector at time $t \in I = [t_0, t_f]$, which is the unique and differentiable solution of the aforementioned initial value problem, and $\boldsymbol{x}_0$ the initial value vector. Moreover, $\boldsymbol{z}(t) \in \mathbb{R}^\kappa$ represents the κ-dimensional independent variables and $\boldsymbol{\theta} \in \mathbb{R}^n$ the parameters of the problem.

In the subsequent sections, we elaborate on how these dynamic systems, described by Eq. (1) are transformed into a cost function to be minimized. Additionally, we outline the steps developed in this study to find the parameters $\boldsymbol{\theta}$ of the ODEs.

2.1 Cost Function

It is further assumed that $\boldsymbol{f}$ is continuously differentiable with respect to the state vector $\boldsymbol{x}$ and the parameters $\boldsymbol{\theta}$, and that there is not any explicit dependence on t and on $\boldsymbol{z}(t)$. Additionally the second term of Eq. (1) depends on the initial conditions, the vector $\boldsymbol{x}$ and the parameters $\boldsymbol{\theta}$. For the minimization process, we employ the Full Information Maximum Likelihood Estimation (FIML) according to the approach of [25, 26] for linear system parameter estimation and the approach of [27] for nonlinear systems. In the case where autoregressive residuals are present, we follow the form of [28, 29, 30]. These works also provide various methodologies for solving this problem. The cost function to be minimized remains the same for both linear and nonlinear cases.

Consider a standard simultaneous difference algebraic equation model of Eq. (1) written as:

$$A\boldsymbol{y}_t - \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{x}_{0,1:s}, \Theta) = \boldsymbol{u}_t \quad (2)$$

with $t = 1 \ldots T$, where $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$ are vectors of observations at time t on the endogenous and determined variables, respectively and $\boldsymbol{u}_t$ is a vector of disturbances at time t. The vector $\boldsymbol{f}$ describes nonlinear functions depending on $\boldsymbol{x}$ and Θ at time t. A is a matrix and Θ is also a matrix of the coefficients to be found. Moreover, we make the following assumptions: A is a nonsingular sxs matrix. Each equation of the system is identified by virtue of the fact that certain elements of A and Θ are known to be equal to zero [30]. In our case (cf. Eq. (2)), A is the identity matrix I and this is identified as the so called reduced form. It is also assumed that $\boldsymbol{u}_t$ follows the multivariate normal distribution. In the case where $\boldsymbol{u}_t$ shows an autoregressive behavior, then it can be further assumed that it is generated by an autoregressive process of the form: $\boldsymbol{u}_t = \mathrm{AR}(n) + \boldsymbol{e}_t$, where $\mathrm{AR}(n)$ is an autoregressive model for $\boldsymbol{u}_t$ of n order without the constant term. The disturbance $\boldsymbol{e}_t$ denotes now the independent and normally distributed residuals with mean zero and the unknown nonsingular variance-covariance matrix Σ, $(\boldsymbol{e}_t \sim NID(0, \Sigma))$. If e.g. $n = 1$ or $n = 2$, then $\boldsymbol{u}_t$ takes according to [28, 29] the following form:

$$\boldsymbol{u}_t = R\boldsymbol{u}_{t-1} + \boldsymbol{e}_t \quad (3)$$

$$\boldsymbol{u}_t = R_1\boldsymbol{u}_{t-1} + R_2\boldsymbol{u}_{t-1} + \boldsymbol{e}_t \quad (4)$$

respectively. R, R_1 and R_2 are matrices with autoregressive coefficient, where some elements of

them may be zero. It is also presumed that the autoregressive order may be different for every equation of the ODE system. Moreover, we can generalize that n can have an order higher than two. These additional autoregressive error terms can be considered as adjustments of the model to the data. Let now be $Y' = [\boldsymbol{y}_1 \dots \boldsymbol{y}_T]$, $X' = [\boldsymbol{x}_1 \dots \boldsymbol{x}_T]$, $U' = [\boldsymbol{u}_1 \dots \boldsymbol{u}_T]$ and $E' = [\boldsymbol{e}_1 \dots \boldsymbol{e}_T]$, then we can write:

$$AY' - F(X, X_{0,1:s}, \Theta) = U' \tag{5}$$

where $(')$ denotes the transpose matrix, Y is a Txs matrix of T observations of each of s dependent (endogenous) variables of the derivatives of X and X_0 is a vector of the initial conditions of X.

Based on these assumptions and the presumption that the density of the residuals is independent from the linearity or nonlinearity of Eq. (1) and (2), the disturbances follow the probability density of the multivariate joint normal distribution and are general of the form $P(E, \Theta|Y) = Q(E, \Theta)$ as described by [25, 27, 29, 31]:

$$\begin{aligned} Q(E, \Theta) &= Q(\boldsymbol{e}_1, \Theta)Q(\boldsymbol{e}_2, \Theta) \dots Q(\boldsymbol{e}_T, \Theta) \\ &= (2\pi)^{-\frac{1}{2}sT}|\Sigma| \exp\left(-\frac{1}{2}\text{tr}(E'E\Sigma^{-1})\right) \end{aligned} \tag{6}$$

where tr denotes the trace and $|\cdot|$ the determinant. Consequently, the full information likelihood functional, which must be maximized, can be expressed as follows:

$$\begin{aligned} L = &-\frac{1}{2}sT\ln(2\pi) + \frac{1}{2}T\ln|A| + \frac{1}{2}T\ln|\Sigma^{-1}| \\ &-\frac{1}{2}\text{tr}(E'E\Sigma^{-1}) \end{aligned} \tag{7}$$

Since the matrix A is the identity matrix, the determinant of the matrix is equal to 1 and the second term thus vanishes. If we assume that Σ is unrestricted, then we can maximize Eq. (7) analytically with respect to Σ [29, 30]:

$$\frac{\partial L}{\partial \Sigma^{-1}} = 0 \tag{8}$$

which implies

$$\Sigma = \frac{E' \cdot E}{T} \tag{9}$$

After substituting this expression into Eq. (7), the concentrated (reduced) likelihood functional takes the following form [27, 29]:

$$L = -\frac{1}{2}sT(\ln(2\pi) + 1) + \frac{1}{2}T\ln|\Sigma^{-1}| \tag{10}$$

The cost function J to be minimized is the opposite of Eq. (10):

$$\begin{aligned} J &= \frac{1}{2}sT(\ln(2\pi) + 1) - \frac{1}{2}T\ln|\Sigma^{-1}| \\ &= \frac{1}{2}sT(\ln(2\pi) + 1) + \frac{1}{2}T\ln|\Sigma| \end{aligned} \tag{11}$$

depending on the determinant of the unknown and non diagonal variance-covariance matrix Σ, which is calculated from the data. If there is only one equation and not an ODE system, then the cost function is reduced to the sum of squared errors.

2.2 Optimizer: Genetic Algorithm

Genetic algorithms (GAs) are adaptive heuristic search algorithms based on the mechanisms of natural selection and genetics. The basic concept of GAs is designed to simulate evolution processes in natural systems, which follow the principles based on the survival of the fittest, first laid down by Charles Darwin. They represent an intelligent exploitation of a random search within a defined search space to solve a problem.

The key points of these algorithms are the reproduction, crossover and mutation, which are performed according to a given probability, just as it happens in real world. Reproduction involves copying (reproducing) solution vectors, crossover includes swapping partial solution vectors and mutation is the process of randomly changing a cell in the string of the solution vector preventing the possibility of the algorithm being trapped. The process continues until the optimizer reaches the optimal solution of the fitness function, which is used to evaluate individuals.

The general steps of a genetic algorithm are the following:

1. definition of the cost function
2. setting the crossover and mutation probabilities
3. random generation of an initial population
4. production of the next generation of the population by probabilistically selecting individuals to produce offsprings via genetic operators e.g. crossover and mutation
5. computation of the cost function for each individual in the current population. Offsprings with better values have higher probability to contribute with one or more offsprings to the next generation, while offsprings with worse values are discarded.
6. repeating the steps 4 and 5 until a relative threshold of accuracy is reached

The reader is referred to [8, 9, 10] for more details about the steps of GAs.

2.3 Framework Description

Before we continue with the several steps of our framework, a question that arises at this point, is how to determine the first term of Eq. (5), i.e. Y. This term is described by functions of predetermined variables X. The derivatives being of first order are written as follows:

$$y_{t,i} = \frac{\Delta x_{t,i}}{\Delta t} = \frac{\left(x_{t+1,i} - x_{t,i}\right)}{\Delta t} = x_{t+1,i} - x_{t,i} \quad (12)$$

or second order:

$$y_{t,i} = \frac{\left(x_{t+1,i} - x_{t-1,i}\right)}{2\Delta t} = \frac{\left(x_{t+1,i} - x_{t-1,i}\right)}{2} \quad (13)$$

for $\Delta t = 1$, $t = 1{:}T$ and $i = 1{:}s$. One way to fit them is to get the lag-1, i.e. time $t-1$, of the variables X in the second term F of Eq. (5) in order to become exogenous and to perform regression fitting the derivatives of first order Y. Another way is to fit the second order Y taking into account the variables X at time t and to perform again regression. However, these two methods do not solve the differential equations, thus they do not yield solutions at time t. Nevertheless, both approaches can be used to obtain an initial approximation of the parameters Θ and to extend the search interval of each parameter.

A different approach is to consider the variables on the second term of Eq. (5) as solutions of the differential equations. Consequently, we integrate the equations for given parameters Θ and initial conditions of X over the entire specified time interval. Since Y in Eq. (5) is centered in time, i.e. defined at $t + 1/2$, we construct a grid with values for X at $t + 1/2$. This means, for given initial conditions and parameters, we have both the term F and a fit of the Y-values resulting from the integration of the ODEs. In this approach and as a first approximation, we don't take into account the errors between the integrated values of the variables, considered as true or latent variables, and the observed values.

Let us now continue with the applied procedure and the basic steps displayed in Fig. 1 and described as follows:

1. **Initialization:** The random initial values of the variables X are very small but not zero since the differential equations of the form (1) do not contain any constant term and the integration of them do not find the trivial solution (zero). The initial values of the parameters Θ are calculated without integration, by performing regression with lagged values of the variables in every equation, as described above.
2. **Integration:** The integration takes place using the Runge-Kutta method of fourth order and a variable time step. As mentioned above, we create a grid with values calculated at $t + 1/2$ and then we compute the unknown variance-covariance matrix Σ. In the first estimation of the parameters Θ, we do not consider any autocorrelation of the residuals. After the first estimation of them, we examine the autocorrelation and partial autocorrelation function of the residuals, we estimate the autocorrelation form of them and we take them into consideration in the next estimations of the parameters. It should be noted that each intermediate estimation of Θ is combined with the next step "Minimization" of this framework and this procedure is repeated until the optimizer converges.
3. **Minimization:** After each integration is completed, we compute the independent residuals and then the cost function J. For the minimization, we use a genetic algorithm using the ga function in Matlab R2016a [32].
4. **Convergence:** The three aforementioned steps are repeated till the genetic algorithm converges, i.e. the change in the parameters Θ between consecutive iterations falls below a specified threshold ϵ, and the optimizer provides the result.

2.4 Data

To estimate the unknown parameters of ODEs, we match the daily data reported by Johns Hopkins in Germany to our simulation results for the period starting on 22/01/2022 and ending on 31/05/2022. Before we begin with the computation of the parameters, we smooth the time series by using the Savitzky-Golay (SG) digital filter [33] on the left side of Eq. 5, i.e., on the differences of active, recovered and total infected persons.

The parameters of the SG digital filter are selected in such a way to ensure that initial values remain zero and do not become negative. Fig. 2 and 3 display the growth and cumulative numbers of total infected, active infected and recovered persons. The total filtered infected persons are determined by the sum of the filtered active infected and recovered persons. Although the given curves are in a very good agreement with the filtered curves in Fig. 3, deviations between them are obtained in Fig. 2, as the noise of the given data is removed. Additionally, Fig. 4 compares the actual data describing the total infected persons with the smoothed data using the SG and the Moving Average (MA) filter [34] with a length of 7. As seen in this figure, the SG curve

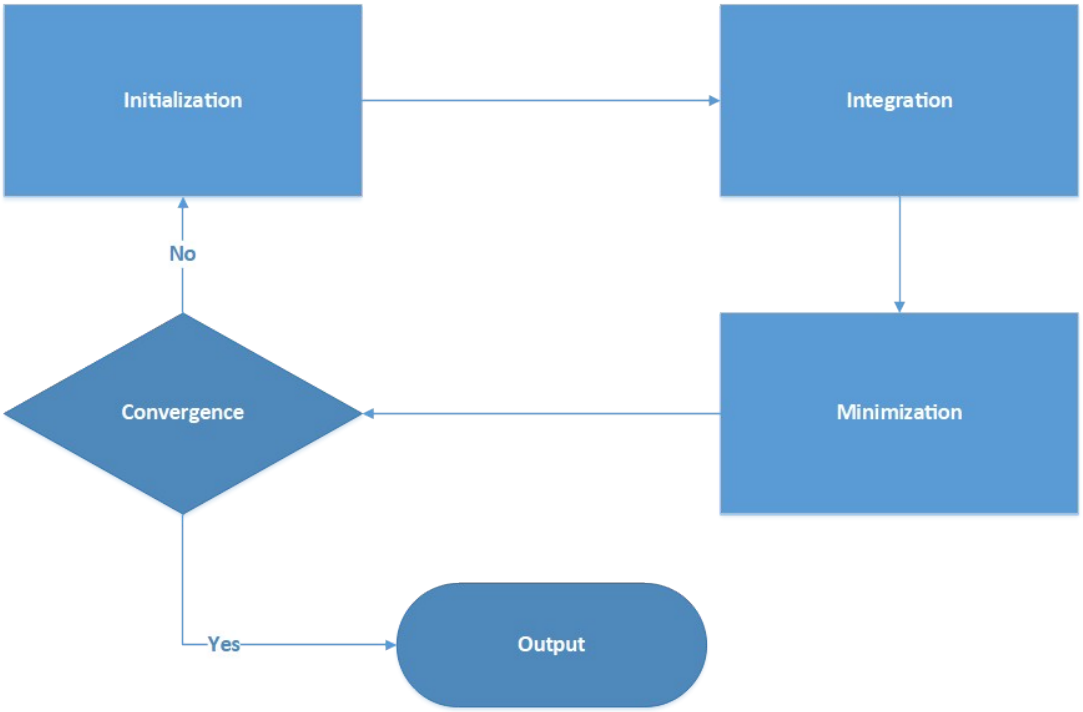

Fig. 1: Main program flow: Description of all framework stages

is slightly smoother than the MA curve, and it is thus considered in the current work in the numerical examples.

3 Numerical Examples

In this section, we apply this procedure in a case that concerns the spread of COVID-19 disease in society, particularly in Germany and we consider two different models for that. The first model is the Bass Model [35] initially applied in marketing but with useful parallels in virus spreading. The second model is the well-known SIR model [36], which stands for susceptible, active infected and recovered individuals.

3.1 The Bass Model

The Bass Model was introduced from Bass in 1969 to describe the launch of a new innovative product in the market. In this work, we apply the Bass model in order to describe the spread of any given disease (in this case COVID-19) in society. The Bass model is a diffusion model and is described by the following diffusion equation:

$$\frac{d\boldsymbol{TI}(t)}{dt} = \boldsymbol{P}(t)(m - \boldsymbol{TI}(t)) \tag{14}$$

where $\boldsymbol{P}(t)$ is the diffusion coefficient, m is the potential number of adopters and $\boldsymbol{TI}(t)$ is the number of adopters at time t. According to [35]: *"The probability that an initial purchase will be made at t given that no purchase has yet been made is a linear function of the number of previous adopters (byers)"*. In our case: *"The probability that an initial infection will be occurred at t given that no infection has yet been occurred is a linear function of the number of previous infections"*. *"No infection or no purchase has yet been occurred"*, means that every individual is infected, or purchased a product for the first time without reinfection or repurchase a product. In this case, $\boldsymbol{P}(t)$ is the diffusion coefficient, m is the potential number of total infected individuals in society and $\boldsymbol{TI}(t)$ is the number of total infected individuals at time t. In this model, there is no distinction between actively infected or recovered. By setting $\boldsymbol{P}(t) = p + \frac{q}{m}\boldsymbol{TI}(t)$, we get the following diffusion equation:

$$\begin{aligned}\frac{d\boldsymbol{TI}(t)}{dt} &= F(\boldsymbol{TI}, \boldsymbol{\Theta}) \\ &= pm + (q-p)\boldsymbol{TI}(t)) - \frac{q}{m}\boldsymbol{TI}(t)^2 \\ &= \theta_1 + \theta_2 x_1 + \theta_3 x_2\end{aligned} \tag{15}$$

Alternatively, in the form of Eq. (5):

$$A = 1, \quad \boldsymbol{\Theta}' = -[\theta_1, \theta_2, \theta_3],$$

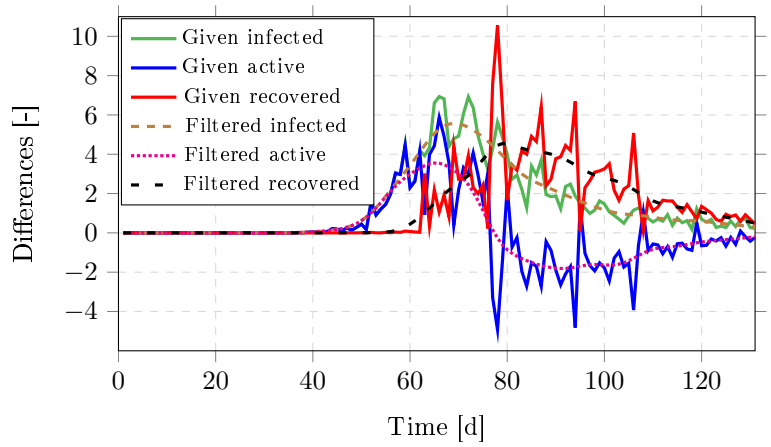

Fig. 2: Differences of total infected, active infected and recovered individuals: given vs. filtered (SG filter) numbers

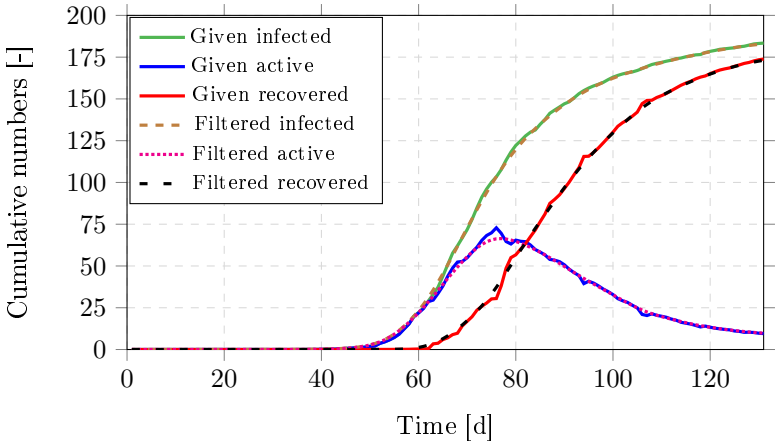

Fig. 3: Cumulative numbers of total infected, active infected and recovered individuals: given vs. filtered (SG filter) numbers

$$\boldsymbol{Y} = \boldsymbol{y} = \frac{d\boldsymbol{TI}(t)}{dt}, \quad X' = [\mathbf{1} \quad \boldsymbol{x}_1 \quad \boldsymbol{x}_2],$$
$$\boldsymbol{F} = \boldsymbol{\Theta}' X, \quad \boldsymbol{U} = \boldsymbol{u},$$

where:

$$\theta_1 = pm, \theta_2 = q - p, \theta_3 = \frac{q}{m},$$
$$\boldsymbol{x}_1 = \boldsymbol{TI}, \boldsymbol{x}_2 = \boldsymbol{TI}^2, \boldsymbol{TI}(0) = 0,$$
$$\boldsymbol{y}(t + \frac{1}{2}) = (\boldsymbol{TI}(t+1) - \boldsymbol{TI}(t)),$$
$$\boldsymbol{x}_1(t + \frac{1}{2}) = \widehat{\boldsymbol{TI}}\left(t + \frac{1}{2}, \boldsymbol{\theta}\right),$$
$$\boldsymbol{x}_2(t + \frac{1}{2}) = \widehat{\boldsymbol{TI}}\left(t + \frac{1}{2}, \boldsymbol{\theta}\right)^2$$

This is a nonlinear differential equation of first order, and even if the analytical solution is known [35], we use the aforementioned procedure for testing purposes. The meaning of the variables and parameters in marketing, as well as the parallelism in spread of a disease are given in Table 1.

According to [35], if we use the lag-1 of the time series on the right hand side of Eq. (15), then we can apply regression to find the parameters $\boldsymbol{\Theta}$. We use these values as the first approximation and then we apply the above-mentioned process for the computation of the parameters $\boldsymbol{\Theta}$ and then of p, q, m from the following relations:

$$m = \frac{-\theta_2 \pm \sqrt{\theta_2^2 - 4\theta_1\theta_3}}{2\theta_1}, p = \frac{\theta_1}{m}, q = p + \theta_2$$

By applying the proposed methodology to the data, we find parameter values, which are presented in Table 2.

Analyzing the autocorrelation and partial autocorrelation graph, we can extract the autoregressive model for the residuals, which is an AR(3) model. After applying this model to the

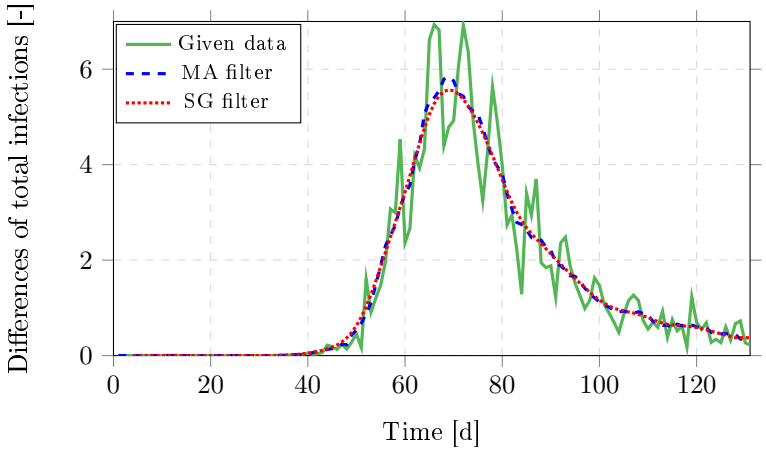

Fig. 4: Differences of total actual (sum of actual active and recovered) infected individuals and their transformations by applying the SG and MA filter (the sum of filtered active and recovered individuals).

Table 1. Explanation of the variables and parameters in the Bass model

Variables, Parameters	Marketing	Spread of Disease
$\boldsymbol{TI}(t)$	cumulative number of the adopters in time t	cumulative number of total infected individuals in time t
m	total potential number of adoptions	total potential number of infected individuals
p	probability of initial purchase (exogenous factor); it describes the innovators in the Bass Model.	probability of the initial infection (exogenous factor), i.e. first outbreak due to travel
$\frac{q}{m}\boldsymbol{TI}(t)$	pressure of the previous adopters to the society, to the non-yet adopters (endogenous factor); imitators in the Bass Model	spread of the disease in the society from interactions between infected and non-infected individuals (endogenous factor)

residuals, we observe a much better behavior of the autocorrelation and partial autocorrelation factors of the remaining residuals. Most of their values lie within the validity interval of the null hypothesis,

Table 2. Estimated parameter values of the Bass model

Parameter	Value
without autoregression	AR(0)
θ_1	$3.39 \cdot 10^{-3}$
θ_2	$1.22 \cdot 10^{-1}$
θ_3	$-6.96 \cdot 10^{-4}$
m	174.89
p	$1.94 \cdot 10^{-5}$
q	$1.22 \cdot 10^{-1}$
with autoregression	AR(3)
θ_1	$1.80 \cdot 10^{-3}$
θ_2	$1.33 \cdot 10^{-1}$
θ_3	$-8.26 \cdot 10^{-4}$
m	160.88
p	$1.12 \cdot 10^{-5}$
q	$1.33 \cdot 10^{-1}$

which are statistically zero.

The values of the parameter p are of the order of 10^{-5}, several order of magnitudes smaller than the values of q. The occurrence of an outbreak is very rare and random. The probability of the initial infection is very small, and the entire phenomenon is evolved from endogenous factors, also from the interaction between infected and non-infected individuals. The maximum number of infected individuals m is smaller than the maximum of the actual values. Also the m value of the AR(3) model is smaller than that of the AR(0) model. Fig. 5 compares the following curves without taking into account the autoregressive residuals:

- The actual data (green curve) after applying the

SG-filter.

- The calculated fit using regression with lags according to the Bass method as described above (blue loosely dashed curve). The solution becomes negative after about 120 days, which is not physically possible. The variation of total infected individuals must be greater than or equal zero.

- The brown densely dotted curve is calculated using a self-consistent integration of the equation with Runge-Kutta method and parameters found from the regression with lagged values of independent variables. There is a significance deviation between the green and the brown curve, indicating that the brown curve can not be the solution of the differential equation.

- The red dashed curve represents the solution derived from the methodology developed in this work and is close to the actual data without facing the problem of negative values as the exogenous regression does.

The corresponding cumulative numbers of the total infected individuals are shown in Fig. 7. The deviation between the green and the brown curve is also here very large, indicating that the parameters found from the regression with lags do not provide a good solution of the differential equation in comparison with the actual data. In Fig. 6 and 8, the solutions of our proposed methodology after applying the AR(3) model for the residuals are presented. The actual and calculated data are in a very good agreement, even though the value of m is smaller than the maximum actual value, which is not physically plausible. The error correction term practically adjusts the model to the data.

3.2 The SIR Model

The first set of dependent variables counts people in each of the groups, each as a function of time. $\boldsymbol{S}(t)$ represents the number of susceptible individuals, $\boldsymbol{I}(t)$ denotes the number of active infected individuals, and $\boldsymbol{R}(t)$ describes the number of recovered individuals plus the deaths from the disease. Obviously, $\boldsymbol{TI}(t) = \boldsymbol{I}(t) + \boldsymbol{R}(t)$ is the total (cumulative) number of infected individuals at time t. The SIR system without the so-called vital dynamics (birth and death), which is a good approximation in a short time of evolution and means $\boldsymbol{S}(t) + \boldsymbol{I}(t) + \boldsymbol{R}(t) = N = \text{const.}$, can be expressed by the following system of ordinary differential equations:

$$\begin{aligned}\frac{d\boldsymbol{S}(t)}{dt} &= -\frac{\beta}{N}\boldsymbol{I}(t)\boldsymbol{S}(t)\\ \frac{d\boldsymbol{I}(t)}{dt} &= \frac{\beta}{N}\boldsymbol{I}(t)\boldsymbol{S}(t) - \gamma\boldsymbol{I}(t)\\ \frac{d\boldsymbol{R}(t)}{dt} &= \gamma\boldsymbol{I}(t)\end{aligned} \tag{16}$$

subject to $\beta, \gamma > 0$, where β is the average number of contacts per person per time and γ is the reciprocal of the average time of an individual to be infectious. The initial values are unknown and they are introduced as parameters in the process. Considering $\boldsymbol{S}(t) = N - \boldsymbol{I}(t) - \boldsymbol{R}(t)$ and $\boldsymbol{TI}(t) = \boldsymbol{I}(t) + \boldsymbol{R}(t)$ in (16), to avoid numerical difficulties during the integration, leads to the following system of differential equations reduced by one equation compared to (16):

$$\begin{aligned}\frac{d\boldsymbol{TI}(t)}{dt} &= \frac{\beta}{N}\boldsymbol{I}(t)(N - \boldsymbol{TI}(t)) = \beta(\boldsymbol{TI}(t) - \boldsymbol{R}(t))\\ &\quad - \frac{\beta}{N}\boldsymbol{TI}(t)(\boldsymbol{TI}(t) - \boldsymbol{R}(t))\\ \frac{d\boldsymbol{R}(t)}{dt} &= \gamma\boldsymbol{I}(t) = \gamma(\boldsymbol{TI}(t) - \boldsymbol{R}(t))\end{aligned} \tag{17}$$

Alternatively, in the form of Eq. (5):

$$A = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \quad \Theta = -\begin{bmatrix}\beta & -\beta/N\\ \gamma & 0\end{bmatrix},$$

$$Y' = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2], \quad X' = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2],$$

$$F = \Theta X, \quad U' = [\boldsymbol{u}_1 \quad \boldsymbol{u}_2],$$

where:

$$\begin{aligned}\boldsymbol{y}_1 &= \frac{d\boldsymbol{TI}}{dt}, \quad \boldsymbol{y}_2 = \frac{d\boldsymbol{R}}{dt},\\ \boldsymbol{x}_1 &= (\boldsymbol{TI} - \boldsymbol{R}), \quad \boldsymbol{x}_2 = \boldsymbol{TI}(\boldsymbol{TI} - \boldsymbol{R})\\ \boldsymbol{y}_1(t+\frac{1}{2}) &= (\boldsymbol{TI}(t+1) - \boldsymbol{TI}(t)),\\ \boldsymbol{y}_2(t+\frac{1}{2}) &= (\boldsymbol{R}(t+1) - \boldsymbol{R}(t)),\\ \boldsymbol{x}_1(t+\frac{1}{2}) &= \left(\widehat{\boldsymbol{TI}}\left(t+\frac{1}{2},\Theta\right) - \widehat{\boldsymbol{R}}\left(t+\frac{1}{2},\Theta\right)\right),\\ \boldsymbol{x}_2(t+\frac{1}{2}) &= \left(\widehat{\boldsymbol{TI}}\left(t+\frac{1}{2},\Theta\right) - \widehat{\boldsymbol{R}}\left(t+\frac{1}{2},\Theta\right)\right)\\ &\quad \widehat{\boldsymbol{TI}}\left(t+\frac{1}{2},\Theta\right)\end{aligned} \tag{18}$$

with $\beta, \gamma, N > 0$ and x_i (for $i = 1,2$) represents the integrated values of the variables, which are dependent on Θ. In this approach, we do not set the total susceptible individual number N equal to the total population of the country. Instead, we calculate

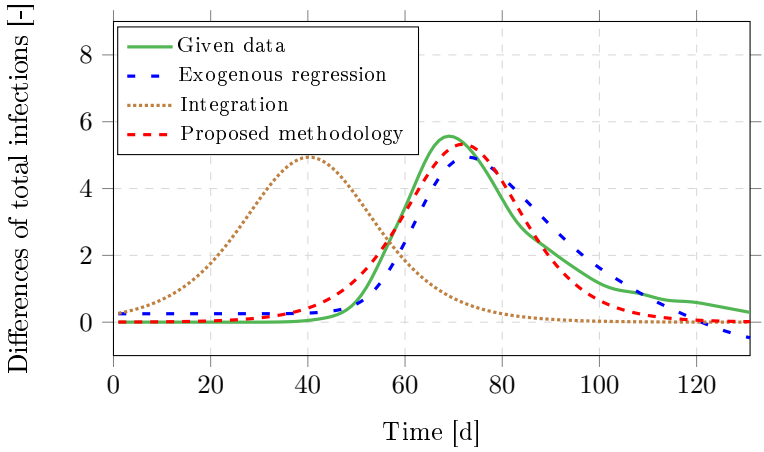

Fig. 5: Differences of total infections: actual infected individuals, differences after the regression with exogenous lag-1 variables, differences after Runge-Kutta integration with parameters found from exogenous regression and differences after applying the proposed methodology

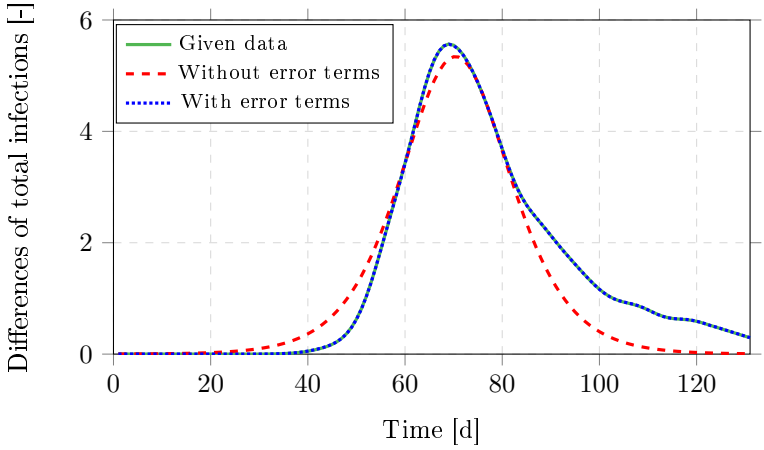

Fig. 6: Differences of total infections: actual infected individuals and differences after applying the proposed methodology without and with the error terms

N from the data. We believe that not all of the people of a country are susceptible because some are immune to the disease, some live in isolation, and others strictly adhere to the political government measures. Additionally, some individuals may be infected without showing symptoms and thus remain outside of the measurement system. In addition, the measurement tests and methods were not very accurate during the outbreak of the unknown disease and there was very little information disseminated from the center of the outbreak. For these reasons, we compute N from the gathered data. Table 3 gives the fitting parameters according to the proposed methodology taken into account autoregressive and non-autoregressive residuals. The calculated number of susceptible individuals N during the first wave is of the order of 194,000. $\boldsymbol{TI}(0)$ and $\boldsymbol{R}(0)$ are the initial conditions for the variables found from the GA as parameters in order to avoid the trivial solution.

Based on the autocorrelation and partial autocorrelation graph, we can identify an autoregressive model AR(3,2) for the residuals. However, in practice, an autoregressive model AR(1,1) is sufficient to reproduce the actual curves because of the very small remaining residuals even if they show an autoregressive behavior. In Fig. 9, 10 and 11, 12, the actual growth and cumulative numbers – total infected (green curves), active infected (blue curves) and recovered (red curves) – are displayed along with the corresponding solutions of the SIR system (dashed, densely dotted and loosely dashed curves) without and with the autoregressive residuals, respectively. The actual data and the corresponding solution of the SIR system are very close to each

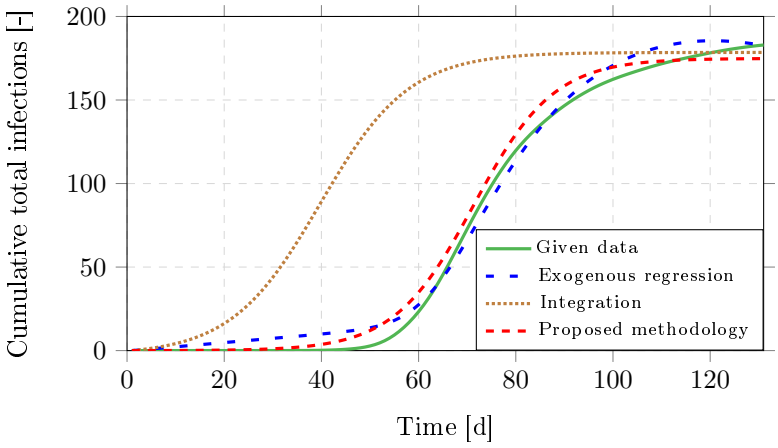

Fig. 7: Cumulative total infections: actual infected individuals, infections after the regression with exogenous lag-1 variables, after the Runge-Kutta integration with parameters found from exogenous regression and after applying the proposed methodology

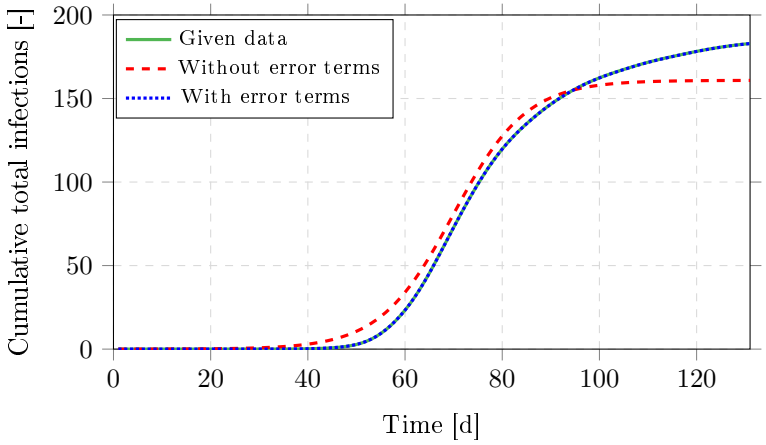

Fig. 8: Cumulative total infections: actual infected individuals and individuals after applying the proposed methodology without and with the error terms

other and by accounting for the error terms, the solution practical represents exactly the actual data. The reproduction of the actual data would be better if the measurement system had worked effectively during the pandemic outbreak of the disease. The mean basic reproduction number $R_0 = \beta/\gamma$ for the entire first wave in Germany is equal to 3.64 for the solution without error terms and 3.33 for the solution with error terms, which are close to each other. Epidemiologically, R_0 indicates the number of new infections an infected individual causes during the infectious period in an otherwise susceptible population. [37] has assumed a value of 3 for R_0 and a piecewise constant function of β and thus for R_0 for the first wave in Germany, which is in good agreement with our results. Also [38] has estimated a mean value of 2.9 for R_0 in an interval of 2.4 – 3.4.

4 Conclusion

In this work, we developed a novel methodology that combines numerical analysis, econometrics and a genetic algorithm as an optimizer to determine the parameters of ODEs concerning initial value problems. This methodology starts with the initialization of the parameters and iterates alternating the integration of the equations, producing a fit of the variable derivatives, and the minimization of the FIML functional, building a dynamic process without predetermined parameters.

We presented the proposed methodology in one use case concerning the evolution of the COVID-19 disease in the German society during the first wave.

In the use case considered, we examined two different models: the Bass model, initially developed for marketing purposes, and the SIR-based model for describing the SARS-CoV-2 outbreak in Germany

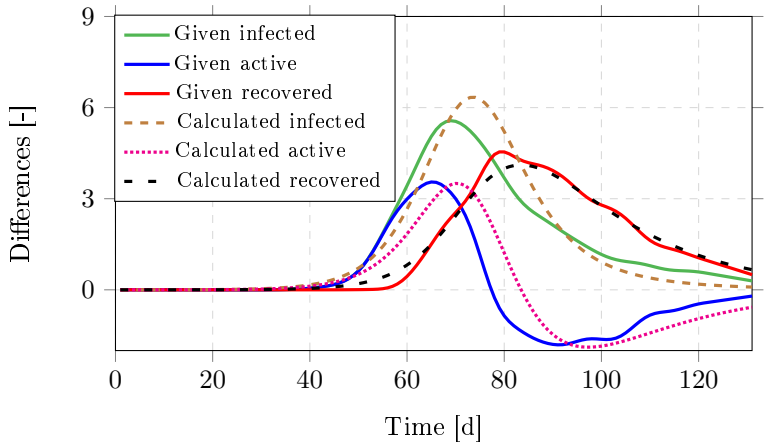

Fig. 9: Differences of total infected, active infected and recovered individuals: given vs. calculated numbers according to the proposed methodology without the error terms

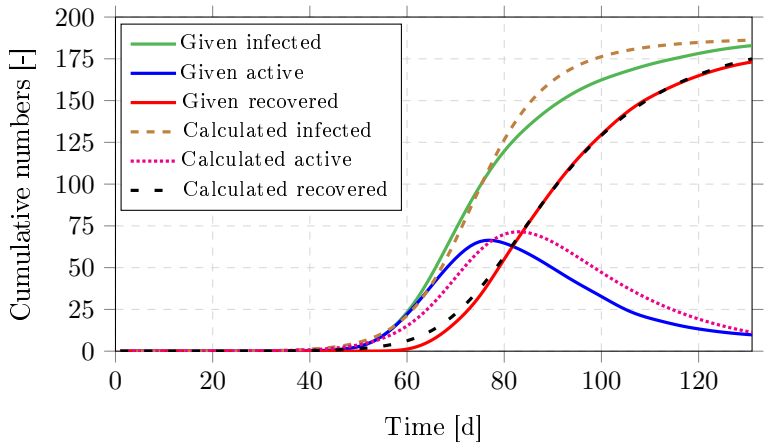

Fig. 10: Cumulative numbers of total infected, active infected and recovered individuals: given vs. calculated numbers according to the proposed methodology without the error terms.

Table 3. Estimated parameter values of the SIR model.

Parameter	**Value**
without autoregression	AR(0,0)
$\boldsymbol{TI}(0)$	$8.78 \cdot 10^{-3}$
$\boldsymbol{R}(0)$	$6.42 \cdot 10^{-3}$
β	$2.09 \cdot 10^{-1}$
γ	$5.75 \cdot 10^{-2}$
N	193.47
with autoregression	AR(3,2)
$\boldsymbol{TI}(0)$	$1.49 \cdot 10^{-2}$
$\boldsymbol{R}(0)$	$8.44 \cdot 10^{-3}$
β	$2.13 \cdot 10^{-1}$
γ	$6.40 \cdot 10^{-2}$
N	187.62

during the first wave. Other effects such as social distancing [37] were modeled by a time-dependent reduction of the transmission rate, but they were not explicitly taken into account. The best results were produced by using autoregression for the residuals appearing in the minimization of the FIML functional in all models and by considering the data published by Johns Hopkins University, which allows for the estimation of unknown model parameters. An appropriate autoregressive model for the residuals is the AR(1,1), which reproduces the real curves with high accuracy. Higher-order autoregressive models can lead to unrealistic values for the susceptible population, demonstrating that the choice of the autoregressive model has a significant impact on the solution of the differential equations.

In future work, we aim to generalize our method for more than two equations. The SEIRD model will be applied by adding the time evolution of

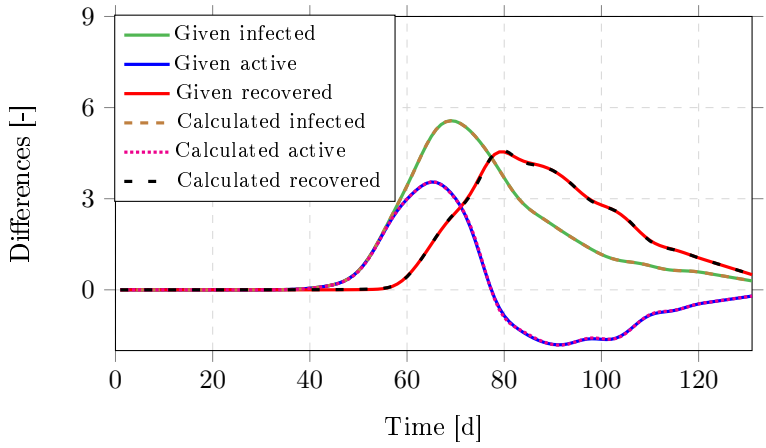

Fig. 11: Differences of total infected, active infected and recovered individuals: given vs. calculated numbers according to the proposed methodology with the error terms

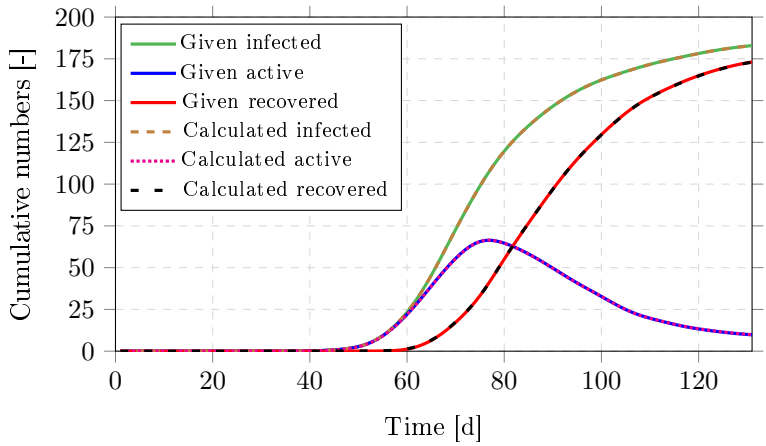

Fig. 12: Cumulative numbers of total infected, active infected and recovered individuals: given vs. calculated numbers according to the proposed methodology with the error term

exposed and deceased individuals with two additional differential equations, also taking into account the vaccination. Moreover, the proposed methodology will be applied to the telecom development, where the evolution of mobile technologies (2G, 3G, 4G, and 5G) over time and their penetration in the Greek market will be examined by applying a competitive initial value model consisting of four nonlinear differential equations.

References:

[1] H. G. Bock, E. Kostina, J. P. Schlöder, Numerical Methods for Parameter Estimation in Nonlinear Differential Algebraic Equations, GAMM-Mitteilungen 30 (2) (2007) 376–408.

[2] M. M. Ali, C. Storey, A. Torn, Application of Stochastic Global Optimization Algorithms to Practical Problems, Journal of Optimization, Theory and Applications 95 (3) (1997) 545–563.

[3] Z. B. Zabinsky, R. L. Smith, Pure adaptive search in global optimization, Mathematical Programming 53 (1992) 323–338.

[4] J. Banga, W. Seide, Global Optimization of Chemical Processes using Stochastic Algorithms, in: C. Floudas, P. Pardalos (Eds.), State of the Art in Global Optimization. Nonconvex Optimization and Its Applications, Springer, Boston, MA, 1996, p. 563–583.

[5] A. Törn, M. Ali, S. Viitanen, Stochastic Global Optimization: Problem Classes and Solution Techniques, Journal of Global Optimization 14 (1999) 437–447.

[6] A. R. Kan, G. Timmer, Stochastic global optimization methods part i: Clustering methods, Mathematical Programming 39 (1987) 27–56.

[7] J. H. Holland, Genetic Algorithms, Scientific American 267 (1992) 66–73.

[8] C. Michalakelis, T. Sphicopoulos, D. Varoutas, Modeling Competition in the Telecommunications Market Based on Concepts of Population Biology, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 41 (2) (2011) 200–210.

[9] C. Michalakelis, C. Christodoulos, D. Varoutas, T. Sphicopoulos, Dynamic estimation of markets exhibiting a prey–predator behavior, Expert Systems with Applications 39 (9) (2012) 7690–7700.

[10] Z. Qiu, Y. Sun, X. He, J. Wei, R. Zhou, J. Bai, S. Du, Application of genetic algorithm combined with improved SEIR model in predicting the epidemic trend of COVID-19, China, Scientific Reports 12 (2022) 1–9.

[11] J. R. Banga, E. Balsa-Canto, C. G. Moles, A. A. Alonso, Improving food processing using modern optimization methods, Trends in Food Science Technology 14 (4) (2003) 131–144.

[12] J. H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, 1975.

[13] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.

[14] N. Andrei, Continuous Nonlinear Optimization for Engineering Applications in GAMS Technology, Vol. 121 of Springer Optimization and Its Applications, 2017.

[15] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Springer, 1993.

[16] P. Pfeifer, J. Timmer, Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting, IET System Biology 1 (2) (2007) 78–88.

[17] H. Bock, Recent advances in parameter identification techniques for ordinary differential equations, in: P. Deuflhard, E. Hairer (Eds.), Numerical Treatment of Inverse Problems in Differential and Integral Equations, 1983, pp. 95–121.

[18] H. G. Bock, Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen, Phd thesis, Universität Bonn, Bonn, Germany (1987).

[19] H. G. Bock, Numerical treatment of inverse problems in chemical reaction kinetics, in: K. Ebert, P. Deuflhard, W. Jäger (Eds.), Modelling of Chemical Reaction Systems, Springer, 1981, pp. 102–125.

[20] K. Menda, L. Laird, M. J. Kochenderfer, R. S. Caceres, Explaining COVID-19 outbreaks with reactive SEIRD models, Scientific Reports 11 (1) (2021) 1–12.

[21] R. Dandekar, G. Barbastathis, Quantifying the effect of quarantine control in covid-19 infectious spread using machine learning, medRxiv (2020).

[22] M. Wieczorek, J. Siłka, M. Woźniak, Neural network powered COVID-19 spread forecasting model, Chaos, Solitons & Fractals 140 (2020) 1–15.

[23] P. Melin, J. C. Monica, D. Sanchez, O. Castillo, Multiple Ensemble Neural Network Models with Fuzzy Response Aggregation for Predicting COVID-19 Time Series: The Case of Mexico, Healthcare 8 (2) (2020) 1–13.

[24] S. E. Holte, A Consistent Direct Method for Estimating Parameters in Ordinary Differential Equations Models, arXiv (2016).

[25] Tjalling, T.C. Koopmans and W.C. Hood, The estimation of simultaneous economic relationships, Wiley, New York, 1953.

[26] G. C. Chow, Two methods of computing full-information maximum likelihood estimates in simultaneous stochastic equations, International Economic Review 9 (1) (1968) 100–112.

[27] Y. Bard, Nonlinear Parameter Estimation, Academic Press, Cambridge, Massachusetts; New York, 1974.

[28] J. D. Sargan, The Maximum Likelihood Estimation of Economic Relationships with Autoregressive Residuals, Econometrica 29 (3) (1961) 414–426.

[29] D. F. Hendry, Maximum Likelihood Estimation of Systems of Simultaneous Regression Equations with Errors Generated by a Vector Autoregressive Process, International Economic Review 12 (2) (1971) 257–272.

[30] J. Durbin, Maximum likelihood estimation of the parameters of a system of simultaneous regression equations, Econometric Theory 4 (1988) 159–170.

[31] T. C. Koopmans (Ed.), Statistical Inference in Dynamic Economic Models, Cowles Commission for Research in Economics, Monograph 10, John Wiley & Sons, Inc., New York, 1950.

[32] I. The MathWorks, MATLAB R2016a, version R2016a (2016).

[33] A. Savitzky, M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Analytical Chemistry 36 (8) (1964) 1627–1639.

[34] P. Prandoni, M. Vetterli, Signal Processing for Communications, Taylor and Francis Group, LLC, USA, 2008.

[35] F. M. Bass, A new product growth for model consumer durables, Management Science 15 (5) (1969) 215–227.

[36] W. O. Kermack, A. G. McKendrick, A Contribution to the Mathematical Theory of Epidemics, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 115 (772) (1927) 700–721.

[37] T. Götz, P. Heidrich, Early stage COVID-19 disease dynamics in Germany: models and parameter identification, Journal of Mathematics in Industry 10 (20) (2020) 1–13.

[38] M. A. Billah, M. M. Miah, M. N. Khan, Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence, PLOS ONE 15 (11) (2020) 1–17.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Both Authors contributed equally to the full study.

Sources of funding for research presented in a scientific article or scientific article itself

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)