Forecasting Models of the Coronavirus (COVID-19) Cumulative Confirmed Cases Using a Hybrid Genetic Programming Method

Konstantinos Salpasaranis and Vasilios Stylianakis

Abstract — The coronavirus disease 2019 (COVID-19) diffusion process, starting in China, caused more than 4600 lives until June 2020 and became a major threat to global public health systems. In Greece, the phenomenon started on February 2020 and it is still being developed. This paper presents the implementation of a hybrid Genetic Programming (hGP) method in finding fitting models of the Coronavirus (COVID 19) for the cumulative confirmed cases in China for the first saturation level until May 2020 and it proposes forecasting models for Greece before summer tourist season. The specific hGP method encapsulates the use of some well-known diffusion models for forecasting purposes, epidemiological models and produces time dependent models with high performance statistical indices. A retrospective study confirmed the excellent forecasting performance of the method until 3 June 2020.

Index Terms — Genetic Programming; Evolutionary algorithms; diffusion models; coronavirus; forecasting.

I. INTRODUCTION

In epidemiology, the research of disease spread as a diffusion process is a subject with a long history. For modeling the diffusion process of infectious diseases, many classical epidemic models have been proposed [1]-[5], such as SIR and their analytical formulations have been proposed [5], [6]. Also, many methods have been implemented on fitting and forecasting the diffusion process phenomenon [8]. A special case of these methods is the diffusion models case [9] which are dynamic and follow a “S” curve over time. In this article the models which have been used are the Gompertz family models, Logistic model, Bass model, BiLogistic and Log-in-Log models [8]-[16] for Coronavirus cumulative confirmed cases. The initial concept was introduced in [9] while in [10] and [11] a hybrid modified GP (hGP) method combined diffusion models and hGP to produce forecasting models.

The Genetic Programming technique (GP) belongs to the Evolutionary Algorithms (EA) subcategory of Computational Intelligence and embeds the biological natural selection idea on an appropriate problem solution [17]-[19].

II. METHODS

A. Analytical Type of SIR Model

An SIR (Susceptible - Infected - Recovered) model is an epidemiological model that calculates the number of infected people with a disease and recovered or died by a disease in a specific population over time. This models’ category involves differential equations relating the number of Susceptible people s(t), the number of Infected people i(t) and number of people who have Recovered r(t). The SIR models such as Kermack & McKendrick model have two different types with vital dynamics and without [1]-[6].

The first SIR model (without vital dynamics) describes an epidemic spread in a short period of time. The second SIR model (with vital dynamics) describes the epidemic spread over a longer period [7], as endemic phenomenon.

The SIR with vital dynamic model [1]-[7] follows:

\[
\frac{ds(t)}{dt} = -β \cdot i(t) \cdot s(t) - μ \cdot s(t) + μ \\
\frac{di(t)}{dt} = β \cdot i(t) \cdot s(t) - μ \cdot i(t) \\
\frac{dr(t)}{dt} = μ \cdot r(t)
\]

where the initial conditions are, i(0) = i0>0, s(0) = s0>0.

Here t denotes the time, β is the infectivity coefficient with β> 0 and μ the recovery-death coefficient, μ>0. Adding Eq. 1 and Eq. 2:

\[
\frac{ds(t)}{dt} + \frac{di(t)}{dt} = \frac{d[s(t)+i(t)]}{dt} = -μ \cdot [s(t) + i(t)] + μ
\]

Solving the first-order ordinary differential equation (ODE) in Eq. 3, applying the u(t) = s(t) + i(t) and \( \frac{du(t)}{dt} = \frac{ds(t)}{dt} + \frac{di(t)}{dt} = \frac{d[s(t)+i(t)]}{dt} \) we get the general solution:

\[
s(t) = c_1 + c_2 \cdot e^{-μt} - i(t)
\]

where c1 and c2 are constants. Now using Eq. 4 in Eq. 2, we get Bernoulli’s Differential Equation (DE):

\[
\frac{di(t)}{dt} = (β \cdot c_1 - μ + β \cdot c_2 \cdot e^{-μt}) \cdot i(t) - β \cdot i^2(t)
\]

The Eq. 5 must be transformed into first-order ODE. So, replacing y(t) = i^-1(t) we get:

\[
\frac{dy(t)}{dt} + (β \cdot c_1 - μ + β \cdot c_2 \cdot e^{-μt}) \cdot y(t) = β
\]

The solution of Eq. 6 becomes:

Published on December 15, 2020.
Konstantinos Salpasaranis, University of Patras, Greece.
(e-mail: salpk@upatras.gr)
Vasilios Stylianakis, University of Patras, Greece.

DOI: http://dx.doi.org/10.24018/ejers.2020.5.12.2129
\[ y(t) = e^{-\int A dt} \cdot [d + \int B \cdot e^{\int A dt} \cdot dt] \quad (7) \]

where parameters \( A = (\beta \cdot c_1 - \mu + \beta \cdot c_2 \cdot e^{-\mu t}) \), \( B = \beta \) and d constant number.

Solving Eq. 7, we get:

\[ y(t) = i^{-1}(t) = e^{-(\beta \cdot c_1 - \mu) \cdot t} \cdot e^{\frac{\beta \cdot c_2}{\mu} \cdot e^{-\mu t}} \cdot [d + \int \beta \cdot e^{(\beta \cdot c_1 - \mu) \cdot t} \cdot e^{\frac{\beta \cdot c_2}{\mu} \cdot e^{-\mu t}} \cdot dt] \quad (8) \]
or

\[ i(t) = \frac{\beta \cdot e^{(\beta \cdot c_1 - \mu) \cdot t} \cdot e^{\frac{\beta \cdot c_2}{\mu} \cdot e^{-\mu t}}}{d + \int \beta \cdot e^{(\beta \cdot c_1 - \mu) \cdot t} \cdot e^{\frac{\beta \cdot c_2}{\mu} \cdot e^{-\mu t}} \cdot dt} \quad (9) \]

The integral \( \int \beta \cdot e^{(\beta \cdot c_1 - \mu) \cdot t} \cdot e^{\frac{\beta \cdot c_2}{\mu} \cdot e^{-\mu t}} dt \) is related to the Exponential Integral function [20]: \( Ei(x) = -\int_{-\infty}^{x} \frac{e^{-t}}{t} dt \) or the equivalent:

\[ Ei(t) = -\int_{-t}^{\infty} \frac{e^{-u}}{u} du \quad (10) \]

where \( u = \frac{\beta \cdot c_2}{\mu} \cdot e^{-\mu t} \Rightarrow du = \beta \cdot c_2 \cdot e^{-\mu t} \cdot dt \).

A simplification of SIR model in Eq. 11 has been implemented in article [6], using Taylor series approximation:

\[ e^{-\mu t} \approx 1 - \mu \cdot t \]

and under consideration

\[ \lim_{t \to \infty} (e^{-\mu t}) = 0 \]

This SIR model in [6] and [21] gives a Logistic type approach:

\[ i(t) = \frac{\lambda}{\beta + \lambda (\frac{\lambda \cdot e^{(\lambda \cdot t - \mu \cdot t)\cdot e^{\frac{\lambda \cdot c_2}{\mu} \cdot e^{-\mu t}}}}{\lambda \cdot t + \mu})} \quad (11) \]

where \( \lambda = \beta - \mu + \beta \cdot c \) and C is a constant.

B. Diffusion Models

In articles [8]-[15], the diffusion phenomenon has been extensively described. The models in this process follows:

\[ \frac{dy(t)}{dt} = [S - y(t)] \cdot f(t) \quad (12) \]

In this equation, \( y(t) \) is the penetration until time \( t \), \( S \) is the level point of saturation and \( f(t) \) is the diffusion's coefficient function [8], [9]. The type of \( f(t) \) function defines the differential equation solution.

1) Gompertz Model

In this article two types of Gompertz models have been implemented and are described with the equations:

\[ y(t) = S \cdot e^{f(t)} \quad (13) \]

\[ y(t) = S \cdot e^{k(t)} + c \quad (14) \]

In (13) and (14), which are called Gompertz I and II, respectively, the \( y(t) \) is the estimated diffusion level at time \( t \), parameter \( S \) is the saturation level and \( f(t) \) is a linear function of time \( t \), such as \( f(t) = a + b \cdot t \). Parameter \( a, b \) and \( c \) are constant numbers [8]-[11], [16].

2) Logistic Model

In general, the Logistic model is given by:

\[ y(t) = \frac{s}{1+e^{f(t)}} \quad (15) \]

In (15), \( y(t) \) represents the penetration of the observed quantity, until time \( t, f(t) \) is a time dependant function, such as \( f(t) = a + b \cdot t \) and \( a, b \) are constants [8]-[10].

3) Richards’ Model

The generalized logistic model, known as Richards’ growth model, introduces an additional constant parameter \( d, d \neq 0 \) to the logistic model [8]:

\[ y(t) = \frac{s}{(1+e^{f(t)})^d} \quad (16) \]

4) Bass Model

The Bass model concept has been implemented in [8]-[12]. In this model, the diffusion phenomenon has two major elements: the diffusion process pioneers (innovators) and those who follow (imitators).

The penetration of the observed quantity \( y(t) \) is represented in (17):

\[ y(t) = \frac{a - c \cdot e^{-bt}}{1 + de^{-bt}} \quad (17) \]

In (5), parameter \( a \) represents the initial crucial number of diffusion adopters. Parameter \( b \) is the addition of \( p \) and \( q \), where \( p \) and \( q \) are the innovators’ and imitators’ coefficients, respectively. Parameter \( c \) equals with \( r \) and \( p \) multiplication, where \( r \) corresponds to a constant number and \( d \) corresponds to \( r \cdot (q/a) \) multiplication [12].

5) Bi-Logistic Model

In many cases, the observed diffusion process has many phases/generations. For this purpose, the Bi-Logistic curve implements the composition of two Logistic curves [13]. So, Eq.6 is:

\[ y(t) = \frac{S_1}{1 + e^{f_1(t)}} + \frac{S_2}{1 + e^{f_2(t)}} \quad (18) \]

where \( f_1(t) = b_1 + c_1 \cdot (t - t_{n1}) \) and \( f_2(t) = b_2 + c_2 \cdot (t - t_{n2}) \). Saturation levels \( S_1 \) and \( S_2 \) are constant numbers.

Parameters \( b_1, c_1, b_2 \) and \( c_2 \) are also constants and \( t_{n1}, t_{n2} \) represents the time for the first and second generation start point, respectively [13], [14]. This and the next model could be useful in the case of epidemic spread waves.

6) Logistic Growing Saturation Level in Logistic Model (LogInLog)

When whole diffusion process follows the Logistic model
and also the saturation level point is time dependent \( S(t) \) and it follows the Logistic model until the upper saturation level point \( S_u \) [13]-[15] as well, then the Eq. 19 depicts the process.

\[
y(t) = \frac{S_u}{1 + e^{f(t)+d}}
\]

(19)

where \( f(t) = b + a \cdot t \) and \( f_u(t) = b_u + a_u \cdot t \) and all the parameters are constants. The dynamic saturation level point follows in Eq. 20

\[
S(t) = \frac{S_u}{1 + g \cdot e^{f_u(t)}}
\]

(20)

where \( g \) parameter is constant number.

This model describes the diffusion process when diffusion process has coated generations, not clearly separated [14], [15].

III. HYBRID GENETIC PROGRAMMING METHOD

A hybrid GP (hGP) has been proposed in articles [9]-[11]. In this GP method the initial population of solutions comprises the optimized diffusion models. In this article epidemiological models such as SIR are additionally introduced in the initial population. The modified hGP consists of three parts, the regression analysis, the genetic algorithm process, and the model selection.

A. Initial Population

In [9], [10] the concept of the randomly created chromosomes in the initial population of the solutions has been adopted. The blocks’ form for the initial population’ construction combines the basic sets of <Constant of S set>, <arithmetic function of FA set>, <variable of M set> and <mathematical function of FM set>. The construction follows specific rules of production.

In the block \( \text{FA} = \{ +, -, *, / \} \), \( \text{FM} = \{ \text{exp, log, ln, cos, sin, Ei, Logistic, Bass, Gompertz I, Gompertz II, BiLogistic, Log-in-Log, SIR} \} \), \( S = \{ 1, 0.5, 0.01, \text{random from mean, min, max of observed data} \} \), \( M = \{ t \} \).

In the initial population, random chosen functions (FA: Arithmetic, FM: Mathematical, diffusion and SIR models), variables, constants, and user defined blocks. The models are optimized with regression analysis [2]. The flowchart of the hGP method is depicted in Fig. 1.

It should be noted that the Exponential Integral function \((\text{Ei})\) in Eq. 10 and SIR model in Eq. 9 and Eq. 11 have been inserted in the function set and in initial population of the hybrid Genetic Programming Method (hGP), respectively.

B. Solutions Representation

The chromosomes/solutions are strings of characters [9]. The Python Programming Language low-level representation of the abstract syntax tree (ast) corresponds to parse trees. For example, the chromosomes \( \text{exp}(2+3 \cdot t) \cdot \text{Richards}(3 \cdot t) \) and \( \text{exp}(2+3 \cdot t)+\text{Exp}(-3 \cdot t) \) are presented in Fig. 2 and in Fig. 3 as strings of characters and parse trees, respectively [9].

C. Evaluation and Fitness Function

The best chromosome/solution selection is performed using Eq. 21 and Eq. 22 for fitting and forecasting purposes, respectively.

\[
SSE = \sum_{t=1}^{T} [\text{raw}(t) - y(t)]^2
\]

(21)

\[
wSSE = \sum_{t=1}^{T} w_t [\text{raw}(t) - y(t)]^2
\]

(22)

In Eq. 21, the Sum of Squared Error (SSE) is calculated
over time $t=1, 2, 3, ..., T$. Also, dataset $\text{raw}(t)$ comprises the observed data over time $t$ and $y(t)$ corresponds to the produced model’s value [10].

In weighted sum of squared error (wSSE) function, the weight factor is $w_i = 1/T, 2/T, 3/T, ... 1$. So, latest data have greater weight than first data of the dataset. In forecasting process, the whole process is divided in subprocesses with a combination of the SSE and wSSE.

Specifically, the whole process is separated into training and forecasting subprocesses. During subprocess of training, the SSE function is implemented as fitness function, while the forecasting estimation is performed with wSSE. A list with sorted solutions by their fitness function’s values is created. Every solution that does not satisfy the precision limit criterion is rejected. The remaining solutions participate as candidates for the Tournament selection and the next operations of the Crossover and/or Mutation [10], [11] are going to be executed.

D. Crossover and Mutation

The crossover operation concerns the selection of two solutions (parents), implementing the Tournament Selection process, in the sorted list. In the specific kind of Tournament selection, several solutions from the list, with the best, good and not so good fitness value, are randomly selected and then the best of those are chosen for the Crossover or Mutation [10], [11].

In each parent characters’ string, one Crossover point is randomly chosen. The substring of each parent which begins at the specific crossover point is interchanged between two parents and two new children are generated (see Fig. 4).

In the mutation process, one string’s point, which corresponds to a function, is randomly chosen. The process replaces the function, with a new random function [10] (see Fig. 5).

\[
\text{Fig. 4. Representation of Crossover in modified hGP as parse trees.}
\]

E. Additional Statistical Indices

Three additional statistical indices have been used in this study. The overall evaluation method uses Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE) and Mean Absolute Error (MAE).

\[
\text{MAPE} = \frac{1}{T} \sum_{t=1}^{T} \frac{\text{raw}(t) - y(t)}{y(t)} \tag{23}
\]

\[
\text{MSE} = \sum_{t=1}^{T} \frac{[\text{raw}(t) - y(t)]^2}{T} \tag{24}
\]

\[
\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |\text{raw}(t) - y(t)| \tag{25}
\]

In Eq. 23, Eq. 24 and Eq. 25, $y(t)$ corresponds to the model’s value for time $t=1, 2, 3, ...$. Also, $\text{raw}(t)$ are the data of the dataset.

F. Datasets

In this study, the hGP method has been implemented on World Health Organization (WHO) datasets [22]. The first dataset presents the cumulative confirmed cases in China. Specifically, the data is derived from the WHO [22], which presents the cumulative confirmed cases in China, from 1 November 2019 until 16 April 2020. The dataset concerns the daily cumulative confirmed cases per 1000 inhabitants and the dataset is comprised by 98 data points.

The second dataset presents the cumulative confirmed cases in Greece. The data is derived from the WHO [22], which presents the cumulative confirmed cases in Greece, from 26 February 2020 until 5 May 2020. The dataset concerns the daily cumulative confirmed cases per 1000 inhabitants and the dataset is comprised by 70 data points.

IV. RESULTS AND DISCUSSION

In this study the SSE is the statistic indicator that has been used for the fitting process and wSSE for forecasting purposes. In this section, the fitting performance of a generated hGP model and Diffusion models are presented.

A. Fitting Performance of the hGP Model for China

Table.1 contains the program execution parameters of the GP method concerning the WHO Coronavirus (COVID-19) data sets in China.

| TABLE 1: PROGRAM FITTING PARAMETERS OF MODIFIED-GP |
|-----------------------------------------------|
| Fitting | Maximum Number of Generations | Evaluation Function | Precision coefficient for candidates |
| value | 200 | SSE | 5$^a$ |

$^a$ Upper limit of the precision for candidates in Crossover and Mutation (Precision coefficient $\times 1000$).
The fitting performance of the first modified hGP model for the Coronavirus (COVID-19) cumulative confirmed cases with the best fitness value (SSE), is presented in Fig. 6.

The y-axis presents the cumulative confirmed cases per 1000 inhabitants and x-axis the day number from the beginning of the phenomenon.

The relative statistical indices SSE, MAPE, MSE, and MAE of the modified-hGP models are presented in Table 2.

TABLE 2: STATISTICAL INDICES FOR hGP FITTING MODEL FOR CHINA

| Fitting     | Statistical Indices for modified GP model |
|-------------|------------------------------------------|
| hGP model   | SSE          | MAE      | MSE      | MAE      |
|             | 4.231776    | 0.113093 | 0.043181 | 0.15714  |

The fitting hGP model follows:

\[
(0.8635185292422792\times(-0.06420695738496063+\text{Exp}(79.46427824949745-
\text{LoginLog}(0.9668182464791351\times(-33.63777176884596))))+(-38.81721623677958-
\text{LoginLog}(11.6868729917097966\times(-
(31.46909708507185))))+(-39.22816461758648-
\text{LoginLog}(12.145070923436792+\text{Exp}(3.2657713319291037-
\text{LoginLog}(\text{LoginLog}(1.8262851168463925\times(-
(22.13788177551026))))+7.389041017798707-0.03421895581029144)))))
\]

As one can see, this method combines different mathematical functions, like Log-in-Log and Exponential, with the independent variable of time. In Table 2, the modified-GP method achieves excellent statistical performance, as it shows SSE value of 4.2317763902.

Someone can say that in the inner format of the model exists the multiple generations’ model of Log-In-Log. This is an indicator of the “nature” for the COVID-19 cumulative cases curve.

This behavior of the hGP program execution is partially explained with the findings of diffusion models’ fitting performance in the next section.

B. Fitting Performance of the Diffusion Models for China

The Fitting Performance of the Diffusion Models is depicted in Fig. 7.

The y-axis presents the cumulative confirmed cases per 1000 inhabitants and x-axis the day numbers from the beginning of the phenomenon.

The relative statistical indices SSE, MAPE, MSE, and MAE of the diffusion models are presented in Table 3.

TABLE 3: STATISTICAL INDICES FOR THE FITTING PERFORMANCE OF DIFFUSION-SIR MODELS FOR CHINA

| Fitting        | Statistical Indices for Diffusion-SIR models |
|----------------|---------------------------------------------|
| Logistic       | SSE          | MAE      | MSE      | MAE      |
| 241.02852      | 0.70147     | 2.45947  | 1.15361  |
| GompertzI      | 404.77145   | 0.20728  | 4.13032  | 1.32454  |
| GompertzII     | 384.84027   | 2.20785  | 3.92694  | 1.36463  |
| Bass           | 7955.4108   | 25.5161  | 81.17767 | 7.52300  |
| Log-In-Log     | 239.20229   | 0.47905  | 2.44083  | 1.11657  |
| BiLogistic     | 241.02852   | 0.70147  | 2.45947  | 1.15361  |
| SIR model      | 326.68684   | 4.67667  | 3.33353  | 1.25005  |
| Richards       | 240.73796   | 0.62816  | 2.45651  | 1.140313 |

In Table 3 the Log-In-Log diffusion model achieves the best statistical performance, with a value of 239.2022973 for SSE and Richards, Logistic and Bi-Logistic follow with better SSE than the other models.

Comparing the statistical indices for Table 2 and Table 3 we conclude that the fitting hGP model achieves the best SSE, MAPE, MSE and MAE values. The relative statistical indices for the diffusion and SIR models in Table 3 have worse values than hGP model.

This is a first indication that the method of hGP can lead us to useful assumptions about the COVID-19 penetration.

C. Proposed Forecasting Models for Cumulative Confirmed Cases in Greece

In this section, forecasting models for the Coronavirus (COVID-19) cumulative confirmed cases in Greece derived by the hybrid Genetic Programming (hGP) method are proposed. The observed period concerns daily data, 70 data points, from the 26 February until 5 May 2020.

In forecasting process, we used diffusion models, SIR model and fitting hGP models derived from COVID-19 cumulative confirmed cases in China. The phenomenon in China preceded more than one month in relation to that in Greece and it has reached the “temporary” saturation level. The diffusion models and hGP models are inserted in the
initial population of hGP and the population are being evaluated with the weighed SSE (wSSE) performance indicator.

Table 4 contains the program execution parameters of the GP method concerning the WHO Coronavirus (COVID-19) data sets in Greece.

|  | Program Parameters of modified GP |
|---|---|
| Fitting Maximum Number of Generations | Evaluation Function | Precision coefficient for candidates |
| value 200 | wSSE | 5* |

* Upper limit of the precision for candidates solutions in Crossover and Mutation (Precision coefficient*1000).

The first hGP model corresponds to the optimistic scenario. In the specific scenario, the forecasting model predicts a value of 2790 cumulative confirmed cases in first days of June 2020 (datapoint 99, 3 June 2020). As one can see the model is a combination of Bi-Logistic part and a linear function’s part. The first forecasting hGP model has the following format:

\[
\frac{2.2285066091270123/(1.0656151198157326+\exp(-2.065099746680325-0.1502436298033256*(t-(48.5436271109599))))} {+ 0.0058907659155309185*(t-(9.281063892171199))} + \frac{0.0270218882361208/(0.16029468427112187 + \exp(-59.53190921241808-1.2505949209299758*(t-103.97484040340785)))}{(0.0270218882361208/(0.16029468427112187 + \exp(-59.53190921241808-1.2505949209299758*(t-103.97484040340785))) + 0.0058907659155309185*(t-(9.281063892171199))}
\]

The model consists of the two mathematical parts, Bi-logistic and linear function of time. The addition of these parts produces the first forecasting model.

The first part, the Bi-Logistic is:

\[
\frac{2.2285066091270123/(1.0656151198157326+\exp(-2.065099746680325-0.1502436298033256*(t-(48.5436271109599))))} {+ 0.0058907659155309185*(t-(9.281063892171199))} + \frac{0.0270218882361208/(0.16029468427112187 + \exp(-59.53190921241808-1.2505949209299758*(t-103.97484040340785)))}{(0.0270218882361208/(0.16029468427112187 + \exp(-59.53190921241808-1.2505949209299758*(t-103.97484040340785))) + 0.0058907659155309185*(t-(9.281063892171199))}
\]

The linear part is:

\[
0.0058907659155309185*(t-(9.281063892171199))
\]

Specifically, this Linear part could generate various forecasting scenarios. The model has the behavior below:

hGP Forecasting model 1 = Bi-Logistic part + Linear part

Fig. 8 depicts an analysis of the hGP Forecasting model 1 and one can see the strong influence of Bi-Logistic part on the specific model.

The second forecasting model predicts a value of 3017 cumulative confirmed cases in first days of June 2020 (datapoint 99, 3 June 2020). As one can see the model is a combination of Logistic part, SIR like part and a Richards’ part. The second forecasting hGP model has the format:

\[
0.472808946385381/(1.44568790088004+0.602712607718828*\exp(2.29589184234964-0.0435044799792727/(t-0.00767935215741175+0.00769684332365882*\exp(6.2003241741478+0.3160838487443237)))))^{0.207804119375963}
\]

This model combines three mathematical parts, Logistic, SIR and Richards. The multiplication of these parts produces the second forecasting model.

Specifically, the Logistic part is:

\[
1/(1.44568790088004+0.602712607718828*\exp(2.29589184234964-0.0435044799792727(t)))
\]

The combination of Richards with SIR is:

\[
0.472808946385381/((-0.00767935215741175+0.00769684332365882*\exp(6.2003241741478-0.3160838487443237(t))))^{0.207804119375963}
\]

Specifically, the SIR-Richards part could produce various forecasting scenarios. The model has the following behavior:

hGP Forecasting model 2 = Logistic part * SIR-Richards part

Fig. 9 depicts an analysis of the second forecasting hGP model. In this model’s behavior, one can see the strong influence of Richards and SIR parts.
Fig. 10 depicts the forecasting performance of hGP models and the first two diffusion models, Bi-Logistic and Richards, with their fitting performance (SSE) and statistical index wSSE. The y-axis presents the cumulative confirmed cases per 1000 inhabitants and x-axis the day number from the beginning of the phenomenon.

Bi-Logistic and Richards models predict the values of 2613 and 2706 cumulative confirmed cases in first days of June 2020 (point 99, 3 June 2020), respectively.

The relative statistical indices wSSE, MAPE, MSE, and MAE of the forecasting hGP models are presented in Table 5.

| Forecasting | Statistical Indices for modified GP model |
|-------------|------------------------------------------|
| hGP Forecasting model 1 | wSSE | MAPE | MSE | MAE |
|             | 0.03039 | 0.99861 | 0.00105 | 0.02612 |
| hGP Forecasting model 2 | 0.03793 | 0.21041 | 0.00103 | 0.02400 |

The statistical indices for the hGP forecasting models have similar performance. The second forecasting hGP model achieves better MAPE which is one of the best-known indices in forecasting. This leads us to assume that the hGP Forecasting model 2 will have better forecasting performance than the hGP Forecasting model 1.

This behavior of the hGP forecasting models could be explained with the findings of diffusion models’ fitting performance in the next section. There is a strong influence of the Bi-Logistic in the forecasting performance, as well as a Richards and SIR model influence.

**D. Fitting Performance of the Diffusion Models for Greece**

The Fitting Performance of the Diffusion Models is depicted in Fig. 11 (using the hGP runtime environment).

The relative statistical indices wSSE, MAPE, MSE, and MAE of the diffusion models are presented in Table 6.

| Fitting            | Statistical Indices for modified GP model |
|--------------------|------------------------------------------|
| Logistic           | wSSE | SSE | MAPE | MSE | MAE   |
| GompertzI          | 0.07214 | 0.14205 | 0.14205 | 0.00203 | 0.03425 |
| GompertzII         | 0.06942 | 0.13677 | 0.76863 | 0.00195 | 0.03445 |
| Bass               | 0.70925 | 1.58685 | 8.55542 | 0.02267 | 0.13223 |
| Log-In-Log         | 0.83984 | 1.40660 | 0.23676 | 0.02009 | 0.11884 |
| BiLogistic         | 0.03527 | 0.08699 | 0.98992 | 0.00124 | 0.02947 |
| SIR model          | 0.06851 | 0.12991 | 0.10769 | 0.00185 | 0.03216 |
| Richards           | 0.06741 | 0.12114 | 0.32080 | 0.00173 | 0.03139 |

In Table 6 the Bi-Logistic achieves the best statistical performance, as it achieves SSE value of 0.086998679. Richards and SIR models follow with worse SSE values than Bi-Logistic. This partially explains the structure of the forecasting models in the previous section C, where the hGP forecasting models consisted of Bi-Logistic, Richards and SIR parts.

**E. The retrospective study of the forecasting performance in Greece**

In this section, we study the forecasting scenarios for Greece retrospectively based on World Health Organization (WHO) datasets [23] for the last forecasted datapoint. Specifically, the retrospective study is derived from the WHO [23], which presents the cumulative confirmed cases in Greece until 3 June 2020. The retrospective dataset concerns the daily cumulative confirmed cases per 1000 inhabitants and the last datapoint (3 June 2020) of cumulative confirmed cases in Greece is 2,918.
The behavior of the hGP forecasting models is explained with the findings of forecasting performance in Table 7. This table presents the absolute error of the hGP forecasting scenarios for the last datapoint.

| TABLE 7: STATISTICAL INDICES FOR HGP FORECASTING MODELS |
|--------------------------------------------------------|
| Forecasting Statistics Indices for modified gp models   |
| Date: June 2020 (3 June 2020)                          |
| HGP Forecasting model 1                                |
| Real Value                                            |
| Forecasted Value                                      |
| Absolute Error                                        |
| Percentage Error (%)                                  |
| 2,790                                                 |
| 127                                                   |
| 4.35                                                  |
| 2,917                                                 |
| 3,017                                                 |
| 100                                                   |
| 3.43                                                  |
| 2,613                                                 |
| 304                                                   |
| 10.42                                                 |
| 2,706                                                 |
| 211                                                   |
| 7.23                                                  |
| HGP Forecasting model 2                                |
| Bi-Logistic model                                     |
| Richards model                                        |

The hGP Forecasting model 2 achieves the best forecasting performance with second the hGP Forecasting model 1. The Bi-Logistic and Richards models predict the worst values for the cumulative confirmed cases in first days of June 2020 (point 99, 3 June 2020), respectively. Also, the Absolute and Percentage Errors of the hGP forecasting models are smaller than the others.

The assumption of the Section C (Table 5 findings) that the hGP Forecasting model 2 would have better forecasting performance than the hGP Forecasting model 1 has been confirmed.

It should be noted that the dataset for the hGP program execution was 70 datapoints and the forecasting horizon was 29 points (until point 99). So, one can say that the hGP programming method provides us forecasting models using few data compared with the forecast horizon.

V. CONCLUSION

In this paper the hybrid Genetic Programming method (hGP) [9]-[11] has been implemented for World Health Organization Coronavirus (COVID-19) datasets in China and Greece. Using the fitting hGP models derived by the implementation in China’s dataset and some widely used diffusion models, the hGP method achieves to produce forecasting models concerning COVID-19 cumulative confirmed cases in Greece.

The fitting and forecasting performance of the modified hGP are presented and the method presents satisfactory performance with the best statistical indices. The combination of diffusion models with GP method indicates that could be produced some forecasting models for Coronavirus cumulative confirmed cases in Greece. Also, the analysis of the models leads to the formation of prediction’s scenarios.

The hybrid Genetic Programming method consists of a forecasting framework that produces time dependant models and/or causal models for short or long-term forecasting, with more than one variable.

Finally, a retrospective study for the produced hGP models confirmed the assumptions about the forecasting performance of the method and that this method could be used to accurately forecast with few data. Specifically, using statistical indices for fitting (SSE) and forecasting (wSSE, MAPE) with appropriate models into the initial population, the hGP combines and produces complicated forecasting models with high accuracy. So, the key features of hGP have achieved the desired results.

FUNDING

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

ETHICAL CONSIDERATIONS

Ethics approval not required for this type of study.

CONFLICTS OF INTEREST

None of the authors have any conflicts of interest or financial ties to disclose.

REFERENCES

[1] Hethcote, H. (2000). The Mathematics of Infectious Diseases. SIAM Review, 42(4), 599-653. Retrieved April 20, 2020, from www.jstor.org/stable/2653135.
[2] Kermack, W. O. and McKendrick, A. G. (1927). Contributions to the mathematical theory of epidemics, part I. Proceedings of the Royal Society of Edinburgh. Section A. Mathematics. 115:700–721.
[3] Kermack, W. O. and McKendrick, A. G. (1932). Contributions to the mathematical theory of epidemics, ii - the problem of endemicity. Proceedings of the Royal Society of Edinburgh. Section A. Mathematics. 138:55–83.
[4] Kermack, W. O. and McKendrick, A. G. (1933). Contributions to the mathematical theory of epidemics, iii - further studies of the problem of endemicity. Proceedings of the Royal Society of Edinburgh. Section A. Mathematics. 141:94–122.
[5] Hethcote H.W. (1989) Three Basic Epidemiological Models. In: Levin S.A., Hallam T.G., Gross L.J. (eds) Applied Mathematical Ecology. Biomathematics, vol 18. Springer, Berlin, Heidelberg.
[6] Shabbir G., Khan H., Sadiq M., “A note on Exact solution of SIR and SIS epidemic models,” e-prints arXiv:1012.5035, 2010.
[7] Weisstein, Eric W. “SIR Model.” From MathWorld—A Wolfram Web Resource. https://mathworld.wolfram.com/SIRModel.html.
[8] Salpasaranis Konstantinos and Stylianakis Vasilios, “A New Empirical Model for Short-Term Forecasting of the Broadband Penetration: A Short Research in Greece,” Modelling and Simulation in Engineering, vol. 2011, Article ID 798960, 10 pages, 2011. doi:10.1155/2011/798960.
[9] Konstantinos Salpasaranis and Vasilios Stylianakis, “A Hybrid Genetic Programming Method in Optimization and Forecasting: A Case Study of the Broadband Penetration in OECD Countries,” Advances in Operations Research, vol. 2012, Article ID 904797, 32 pages, 2012. doi:10.1155/2012/904797.
[10] Konstantinos Salpasaranis, Vasilios Stylianakis, and Stavros Kotsopoulos, “Combining Diffusion Models and Macroeconomic Indicators with a Modified Genetic Programming Method: Implementation in Forecasting the Number of Mobile Telecommunications Subscribers in OECD Countries,” Advances in Operations Research, vol. 2014, Article ID 568478, 20 pages, 2014. doi:10.1155/2014/568478.
[11] Konstantinos Salpasaranis, Vasilios Stylianakis, “Forecasting the OECD Fixed Broadband Penetration with Genetic Programming Method, diffusion models and macro-economic indicators”, Image Processing & Communications, 2017 http://ipc.upt.edu.pl/index.php/ipc/article/view/65.
[12] Bass F. M., “A new product growth for model consumer durables,” Management Science, vol. 15, no. 5, pp. 215–227, 1969.
[13] P. Meyer, “Bi-logistic growth,” Technological Forecasting and Social Change, vol. 47, no. 1, pp. 89–102, 1994.
[14] P. S. Meyer and J. H. Ausubel, “Carrying capacity: a model with logistically varying limits,” Technological Forecasting and Social Change, vol. 61, no. 3, pp. 209–214, 1999.
[15] M. N. Sharif and K. Ramanathan, “Binomial innovation diffusion models with dynamic potential adopter population,” Technological Forecasting and Social Change, vol. 20, no. 1, pp. 63–87, 1981.
[16] T. Kamalakis, I. Neokosmidis, D. Varoutas, and T. Sphicopoulos, “Demand and price evolution forecasting as tools for facilitating the roadmapping process of the photonic component industry,” in Proceedings of the World Academy Of Science, Engineering And Technology, vol. 18, 2006.
[17] Holland, J. H., “Adaptation in Natural and Artificial Systems”, University of Michigan Press, 1975.
[18] Koza J.R., “Genetic Programming: On the programming of Computers by Means of Natural Selection”, The MIT Press (1992).
[19] Koza J.R., “Genetic programming for economic modeling”, Statistics and Computing 4(2), 187–197 (1994).
[20] Weisstein, Eric W. “Exponential Integral.” From MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/ExponentialIntegral.html.
[21] Weisstein, Eric W. “Gompertz Constant.” From MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/GompertzConstant.html.
[22] World Health Organization (WHO), https://covid19.who.int/region/wpro/country/cn, April 2020.
[23] World Health Organization (WHO), https://covid19.who.int/region/wpro/country/cn, June 2020.