Estimation of COVID-19 epidemic curves using genetic programming algorithm

Nikola Anđelić, Sandi Baressi Šegota, Ivan Lorencin, Vedran Mrzljak and Zlatan Car
University of Rijeka, Faculty of Engineering, Rijeka, Croatia

Abstract
This paper investigates the possibility of the implementation of Genetic Programming (GP) algorithm on a publicly available COVID-19 data set, in order to obtain mathematical models which could be used for estimation of confirmed, deceased, and recovered cases and the estimation of epidemiology curve for specific countries, with a high number of cases, such as China, Italy, Spain, and USA and as well as on the global scale. The conducted investigation shows that the best mathematical models produced for estimating confirmed and deceased cases achieved $R^2$ scores of 0.999, while the models developed for estimation of recovered cases achieved the $R^2$ score of 0.998. The equations generated for confirmed, deceased, and recovered cases were combined in order to estimate the epidemiology curve of specific countries and on the global scale. The estimated epidemiology curve for each country obtained from these equations is almost identical to the real data contained within the data set.

Keywords
COVID-19, disease spread modeling, evolutionary computing, genetic programming, machine learning

Introduction
COVID-19, a novel coronavirus disease represents one of the greatest challenges in modern human history, both from the health and, consequently, from a socio-economic standpoint. COVID-19 is caused by the virus SARS-CoV-2, a member of Betacoronavirus family. Research shows that the worldwide spread of the virus has started in the Chinese province of Wuhan, where the first cases of infection were observed in late 2019. It can be observed that the rapid spread of COVID-19...
caused a considerable strain on public institutions and private companies. Such an impact is particularly noticeable in the healthcare system, especially in places where health professionals are overburdened and there is a lack of needed medical equipment. Such characteristics are particularly emphasized in Europe, especially in Italy and Spain. The spread of viral infections can be described with the epidemic curve, the time-series data of the number of active cases per unit of time, which is also a common practice with COVID-19. Such approach is necessary in order to define the ratio between maximal number of infection cases and the maximal capacity of the country’s healthcare system. In order to reduce this ratio, it is necessary to define the strategy which includes social and health measures against the spread of the infection. Each different applied strategy will result in a specific epidemic curve. In order to determine the optimal strategy for handling the COVID-19 crisis, it is necessary to estimate the potential epidemic curves, according to the chosen strategy and to compare their maximal values with the healthcare capacity. For these reasons, modeling of infectious diseases can be of significant importance.

Methods for estimation of COVID-19 spread are intensively investigated in order to define the optimal method for epidemic curve estimation. Proposed estimations are obtained by using statistical modeling, and did not provide accurate estimations of the epidemic curve. As another approach, artificial intelligence (AI) modeling is proposed. Such approach has been proven as a powerful tool in various branches of science and technology, including clinical practice. There have been attempts to implement AI methods for handling the COVID-19 crisis. McCall offers a view into AI and its use in combating the COVID-19, concluding that it can either be used to actively combat the current and future infections or be used in the spread analyses after the infection has subdued. One of the articles she cites is by Richardson et al. in which the authors demonstrate the successful use of BenevolentAI in determining the potential treatments for COVID-19, determining Baricitinib to be one such medicine. Hu et al. show the use of a modified stacked auto-encoder in prediction of disease spread, with their models achieving good results with prediction of confirmed cases in China. Gozes et al. show the rapid AI development cycle for the use in coronavirus disease pandemic. Through its use, they achieve a relatively high (97.2%) accuracy using CT scans of infected and non-infected patients. Ye et al. demonstrate a CNN trained AI system, using deep convolutional neural networks, which achieves a high AUC score (0.9791). Zeng et al. demonstrate an ANN model of coronavirus spread. While well fitted to existing data, time has shown their predictions to be inaccurate.

One of the approaches based on artificial intelligence is Genetic Programming (GP). Such approach offers a possibility for determination of unique mathematical expression that describes the epidemic curve. The aforementioned expression can be utilized in order to estimate pandemic curves, according to selected COVID-19 strategy. GP has a history of implementation in medicine-based tasks. Tan et al. show a GP approach to oral cancer prognosis. Authors use a small data set of 31 cases. With feature selection of smoking, drinking, tobacco chewing, histological differentiation of SCC and oncogene p63, authors manage to achieve average scores of 83.87% accuracy and AUC score of 0.8341 for the classification task. Ain et al. show the use of Genetic Programming for feature selection and construction for skin cancer image classification. Authors conclude that GP as classification algorithm provides a good solution for distinguishing between benign and malignant cancer images. Ain et al. demonstrate the use of Genetic Programming on the same problem, but through the use of local and global image extracted features. Authors compare six classification algorithms, along with GP and conclude that GP provides better results. D’Angelo et al. propose a solution for distinguishing between bacterial and viral meningitis techniques using genetic programming algorithm in comparison with decision tree algorithm. Authors observe
two cases: one in which algorithms are trained using blood and cerebrospinal parameters, and second in which only cerebrospinal parameters are observed. GP shows good results in both cases, with only a few false positives provided, which given the case can be considered as not dangerous. Senatore et al.²⁴ show the use of cartesian Genetic Programming in handwriting analysis. They use this for classification of patients with Parkinson disease and their approach is evaluated using publicly available PaHaW data set. Results show that models obtained with GP can be applied on this task, with high accuracy.

From above-presented facts, following questions can be raised:

• Is it possible to utilize the GP algorithm to obtain mathematical models that could estimate the number of confirmed, deceased, and recovered cases of countries such as China, Italy, Spain, or the USA?
• Based on the obtained mathematical models for confirmed, deceased, and recovered cases, is it possible to define the symbolic expression of the epidemiology trend for China, Italy, Spain, and the USA?
• Is it possible to utilize the GP algorithm to generate mathematical model that could estimate global epidemiological trend, through the combination of the global mathematical models of deceased, confirmed and recovered cases?

In this paper, the GP algorithm will be utilized in order to develop mathematical models that could estimate the number of confirmed, deceased, and recovered cases, as well as the number of currently active cases—the epidemiology trend of specific countries (China, Italy, Spain, and USA). These countries were selected based on the large number of confirmed, deceased, and recovered cases. The number of confirmed, deceased, and recovered cases will be combined using equation (1) for each date in order to achieve the number of active cases for each date in the data set.

Background and literature review

In this section, a brief overview of literature from the connected previous research will be presented. Overview of modelling infective diseases will be presented first, followed by the description of GP, and finally the comparison of GP with similar AI algorithms.

Modelling the spread of infective diseases

Modelling the spread of infective diseases has been a hot topic of research for many years. It can be easily understood why—as predicting the spread of an infective disease can be crucial in making decisions on how to approach the curbing of its negative effects.²⁵ The increase of the use of data-based disease spread models in health decision making has been growing more apparent lately.²⁶ Traditionally these models were developed using various statistical techniques in an attempt to create a data-based model predicting the spread of disease.²⁷⁻²⁹ Still, shortcomings of such methods are apparent in that the models are fitted manually to the existing data, and many of the seemingly unimportant variables can be ignored, due to researchers failing to note their importance.³⁰ To avoid this issue, AI based modelling can be implemented in the attempt to predict the infectious disease spread.³¹ Certain issues do arise with this approach; main of which are the needs for a large amount of data,³² extensive training times,³³ and the fact that AI provided models can be hard to understand³⁴ and complex to implement.³⁵ Still, the ability of AI to fit to the existing data, without the need for the researcher to take into account all the possible variables, can provide incredibly precise models.³⁶,³⁷ The need for large amount of data is getting to be less of
the problem, with the tendency to collect and store large amounts of data being more and more prevalent in the medicine of today,\textsuperscript{38,39} with training times being addressed with the growing performance of computing machines.\textsuperscript{40} The models being hard to implement, mostly caused due to not being programming language agnostic, and hard to understand can be addressed using novel methods such as Genetic Programming.

\textit{Brief overview of Genetic Programming}

Genetic Programming (GP) is a machine learning algorithm originally used to evolve computer programs using evolutionary computing principles.\textsuperscript{41} When used for regression, GP algorithm is sometimes referred to as Symbolic Regressor, due to the symbolic format of solution it uses and provides.\textsuperscript{42} As this method can also be applied to the mathematical expressions, GP can be used to evolve mathematical expressions.\textsuperscript{43} As mentioned, GP is a machine learning algorithm, which means it uses existing data in order to train and create models. GP randomly creates a number of candidate solutions which make up the population. Then, these solutions are evaluated and best of them recombined in an attempt to find a better solution. This process will be briefly described in this section.

The principle of Genetic Programming is the creation of equations in the form of trees,\textsuperscript{44} which are tested to determine their solution quality. The equations when applied to a data set provide predicted values, which, in comparison to the real values, have a certain error. Those solutions which have a smaller error are considered to be better. The GP algorithm generates the set of initial solutions using fills the trees using “grow” and “full” techniques. “Grow” technique allows for trees with less then maximal depth, while “full” guarantees that the generated trees will reach the specified maximal depth.\textsuperscript{45} The generated trees describe the equation which creates the relation between inputs $X = [x_1, x_2, ..., x_n]$ and the output $Y$.

The question of the need to format the equations as trees within the algorithm can arise. This is done in order to simplify the evolutionary computing (EC) operations—reproduction, crossover, and mutation; which are performed during the process of training the GP models.\textsuperscript{41} The training consists of the following steps\textsuperscript{46}:

- determine the fitness of each individual solution in the solution set,
- apply the EC operations on solutions selected randomly using fitness proportional selection, and
- generate the next population using the results of the EC operations.

It is important to note that the GP algorithm uses fitness proportional selection when selecting candidates for EC operations. Solutions with higher fitness are those solutions which provide a lower error on the training data when applied as the relation between the inputs and the outputs. The idea of GP is that through the recombination and mutation of those particular solutions, which have better fitness compared to others in the current population, the optimal solution can be found.\textsuperscript{47} The brief explanation of EC operations is as follows\textsuperscript{47,48}:

- **reproduction** is performed on a single candidate solution and it copies it into the next population without any modifications,
- **crossover** is performed on two candidate solutions and it works by selecting random node of both trees, separating the candidate solution trees at those vertices and splicing them together while keeping the root node of the first selected tree, and
• **mutation** which randomly modifies a candidate solution tree and copies it into the next population with said modifications.

Another setting of importance are hyperparameters, which are described, along with their values in Section 3.2.2.

**Comparison of Genetic Programming to other algorithms**

Main difference between the GP and similar algorithms is that GP is an algorithm which combines two AI fields—machine learning (ML) and evolutionary computing (EC). EC is commonly used in the fields of solution space search and optimization. 49,50 ML techniques consist of algorithms which tune their internal parameters based on the existing training data. 51 Combinations of these two branches of AI are common, such as using EC algorithms to optimize the hyperparameters of an ANN52 or to directly optimize the neuron connection weights.53 GP differs from such methods by, in its simplest form, being directly based on a genetic algorithm.46 In a sense, GP is a genetic algorithm with a specific shape of the candidate solutions and a fitness function calculated based on the pre-existing data set. A crucial difference between the two is that since the algorithm is applied on the dataset it cannot be directly used for optimization tasks, such as the ones for which EC algorithms are commonly applied, and is instead used for regression and classification tasks.46

Main benefit of the GP algorithm is the earlier mentioned symbolic shape of the models it develops. Models trained with GP are returned as symbolic expressions representing computer programs or mathematical expressions.45 Most models trained using other ML methods, such as ANNs, return models which are hard to understand.54 In addition to this, models created by ANNs are usually specific to the programming language or module used for their creation,55 while the models created by GP are language agnostic. These create a barrier in the wide use of models created with other ML algorithms, as problems can arise due to other researcher’s unfamiliarity with the framework used to create them, along with other issues such as version incompatibilities of modules used. Using models generated in this manner within other existing software packages can be extremely hard, and might require specific gateway software to be created for intercommunication. This is a very apparent issue when ANN models are used for interdisciplinary research, as non-engineering team members can have issues properly utilizing the models created. GP addresses this issue by generating models which can, with minor modifications—in most cases just a simple textual replacement of operation names, be used in a wide range of software packages and programming languages.56

Still, GP has its shortcomings in comparison to ML algorithms. Due to the training process being EC based it can take a significantly higher time compared to algorithms such as ANNs, even further addled by the fact that GP model training cannot be further accelerated using graphical processing units (GPUs) due to the fact that GP does not store information as tensors during the execution.57,58 Additionally, issues such as the aforementioned bloat can cause extremely high memory usage, and stop the models from converging to a quality solution. These issues mean that GP requires significantly more fine tuning when compared to algorithms like ANNs.59,60

**Methodology**

Methodology used within this research will be presented in this section. The description of the used data set is initially given, followed by a brief description of research-specific GP settings and the description of candidate solution quality evaluation.
Data set description

Data set used in this research is publicly available and obtained from the repository by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) which is supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). The data set in question is structured as a time series data set. Data contained shows daily number of confirmed, deceased, and recovered COVID-19 cases in 401 locations, over 78 days, from January 22nd 2020 to April 8th 2020. The date of January 22nd 2020 will be referred to as “start date” in the paper, signifying the starting date in the dataset. This date was used in previous research regarding the modelling of COVID-19 spread and effects.\(^{36,61}\)

To use the data for GP training the data set needs to be reformatted to a regression data set. To achieve this, first each location’s latitude and longitude are written in a row, while the date of the data collection is transformed into an integer representing the number of days elapsed since the first entry in the dataset—22nd January 2020. Finally, the number of confirmed, deceased, and recovered cases is written. This is repeated for each date and location combination yielding three data sets, one for each patient group (confirmed, deceased, and recovered). The data sets are then shuffled to prevent any influence of non-existing patterns on learning,\(^{54}\) and to give a more randomized learning data to algorithm which should result in higher quality models.\(^{51}\) For generating mathematical expressions for confirmed, deceased, and recovered cases in China only the Hubei province was used which means that latitude and longitude are constant and the number of days since the data collection of the cases began is a variable value. For Italy (Milan) and Spain (Madrid) the latitude and longitude are also fixed. For the USA and global model for confirmed, deceased, and recovered cases at each date all data are summed up across localities. In case of USA for each date all localities with reported cases are summed up. The same procedure was used for the development of the global model.

Genetic Programming methodology

This subsection will briefly describe the process of adjusting and tuning the GP algorithm within the presented research.

Genetic Programming implementation. First step in GP use is setting the appropriate values as input and output variables.\(^{54}\) In the presented research the input \(x_1\) marks the geographical latitude, \(x_2\) the geographical longitude and \(x_3\) the number of days elapsed since the start of data collection—22nd of January 2020. The number of cases in patient groups are given as \(y_C\), \(y_D\), and \(y_R\); \(y_C\) representing the number of patients which have been confirmed as infected, \(y_R\) the number of recovered patients, and \(y_D\) the number of deceased patients. Two things are of note here.

The inputs \(x_1\) and \(x_2\) are, as mentioned earlier, set as either a user defined constant or a variable. The GP algorithm itself does not distinguish between the two, and the inputs in question are treated as variables.\(^{46}\) The difference is only apparent when observing the input data set. As the data set subsection is used for local model training, the location latitude and longitude are fixed.

The second thing of note is the fact that three separate outputs exist. GP can only regress a single variable at a time, so three separate models need to be created.\(^{51}\) While it would be possible to create a combined output—in the sense of regressing the epidemiology curve value, that is the number of active cases \(y_A\), which is defined as:

\[
y_A = y_C - (y_D + y_R).
\] (1)
This approach would simplify the training process, as only a single value would need to be regressed—but it would not allow the individual models to be presented, which would mean the loss of potentially important information. The individual numbers of confirmed, deceased, and recovered patients can provide important strategical information in planning the strategy of dealing with the infection.

Hyperparameters of the Genetic Programming algorithm. Determining the exact values of hyperparameters is a complex task, because slight changes of hyperparameters can cause significantly different algorithm behaviors. In this research, instead of a manual selection of the hyperparameters the upper and lower bounds of the hyperparameter value are defined and randomly uniformly selected from within this range. If the obtained model is not precise enough, the new random set of hyperparameters is selected. This process is repeated until a high quality model is generated. The list of hyperparameters used and their ranges selected by authors are given in Table 1 in order that will be used during result presentation, with brief explanation of hyperparameters following below.

‘Population’ describes the amount of candidate solutions to be generated in each iteration of the algorithm—the so-called generation, number of which is defined with the ‘generation’ hyperparameter. “Tournament Size” describes the amount of candidate solutions to be used within the fitness proportionate selection methods. Crossover and mutation probabilities are the probabilities of EC operations briefly described in section 2 occurring, with the probability of reproduction equaling the difference between 1.0 and the sum of listed probabilities. “Maximum Samples” describes the percentage of candidate solutions which are used for the crossover and mutation operations, with the remainder being reproduced in the following generation. “Constant Range” describes the range of constants GP algorithm can pick to use as operands for operations within the solutions. “Initial Tree Depth” defines the minimum and maximum depth of trees in the initial population, while “Stopping Criteria Coefficient” defines the value of error that needs to be achieved in order to stop the execution. If this value is not reached, the execution stops after the number of generations defined via the hyperparameter “generations” is reached. Finally, “Parsimony Coefficient” is a coefficient which is introduced to prevent bloat, which refers to the phenomenon of increase in equation size without the appropriate model quality benefits.

| Hyperparameter                  | Lower bound | Upper bound |
|--------------------------------|-------------|-------------|
| Population                     | 100         | 1000        |
| Generation                     | 100         | 200         |
| Tournament size                | 10          | 50          |
| Crossover probability          | 0.7         | 1.0         |
| Subtree mutation probability   | 0           | 1           |
| Hoist mutation probability     | 0           | 1           |
| Point mutation probability     | 0           | 1           |
| Maximum samples                | 0.9         | 1           |
| Constant range                 | -10000      | 10000       |
| Minimal initial tree depth     | 3           | 6           |
| Maximal initial tree depth     | 8           | 12          |
| Stopping criteria coefficient  | 0.001       | 1.0         |
| Parsimony coefficient          | 0.001       | 0.1         |
In addition to the previously defined hyperparameters the gene set, namely the operation vector—the set of operations to be used within the algorithm, needs to be defined. The operations in this paper are given in Table 2.

With these values defined, the training of models using GP algorithm can start. The next step is deciding on the way of evaluating the candidate solutions, which is described below.

**Candidate solution evaluation**

After the initial training, quality of the obtained solutions need to be evaluated. Mean Absolute Error $MAE$ was used as a fitness value during the training process. Let $y$ be vector containing real data and $\hat{y}$ be the vector containing corresponding data the model obtained from the model, if $m$ is the number of samples in each, then $MAE$ can be calculated with:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|.$$  \hspace{1cm} (2)

While $MAE$ provides a good insight into the model quality it is not necessarily a good metric for further candidate solution evaluation, when the process of evaluation is performed on the training set. For this reason, coefficient of determination ($R^2$) is used. $R^2$ calculates a value which ranges from 0 to 1 ($R^2 \in [0, 1]$). It compares two sets of solutions—the real data $y$ and the data obtained by the model $\hat{y}$, in terms of variance. $R^2$ provides the answer to the amount of variance contained inside the data $y$, which is explained in the data $\hat{y}$ as a model output. In other words, $R^2$ value of 1.0 means that no variance is left unexplained between two sets, while the value of 0.0 means none of the variance contained in real data $y$ is explained in the model data $\hat{y}$. $R^2$ is calculated as a factor of residual and total variance per:

$$R^2 = 1 - \frac{S_{\text{RESIDUAL}}}{S_{\text{TOTAL}}} = 1 - \frac{\sum_{i=0}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{m}(y_i - \frac{1}{m}\sum_{i=0}^{m}y_i)^2},$$ \hspace{1cm} (3)

with $m$ being the length of vectors $\hat{y}$ and $y$—the number of samples in the testing data set.

**Results and discussion**

Since the COVID-19 pandemic began; each country responded with different measures at different time periods so it was impossible to develop the equation which could globally predict the COVID-19 epidemiology trend using latitude, longitude and outbreak period as an input variable. Instead the global model is made with the number of days since the start of the data collection as an input variable and number of patients in each group as the output variable. In order to develop such model, the data set was modified in terms of summing the number of cases, across locations, for each day. In addition to the global model total of four different models were created and these are China model, Italian model, Spanish model and USA model. These four models are interesting and should be studied.

| Arity | Operation |
|-------|-----------|
| 2     | Addition, subtraction, multiplication, division |
| 1     | Square root, absolute value, natural logarithm, sine, cosine, tangent, negation |
| N     | Maximum of values, minimal of values |

Table 2. The operations used to generate the candidate solutions.
further since each of these countries reacted to the outbreak differently. Among them the China model is the most reliable since the number of confirmed and deceased patients is decreasing while the number of recovered patients is growing. However, the Italian, Spanish and USA models are still unpredictable since each day many new instances of confirmed and deceased cases appear while the number of recovered patients remains small when compared to the previous two groups. In each case the produced mathematical equation for prediction of epidemiology trend consists of three equations which were obtained for recorded confirmed, deceased, and recovered cases. The mathematical equation in general form for prediction of epidemiology trend is given in equation (1).

**China model**

For China model the province of Hubei was chosen as a training/testing model based on the number of confirmed cases, as well as the number of deceased and recovered cases. The equations for confirmed, deceased, and recovered patients are shown in Table 3 with $R^2$ scores, while the performance comparison of each equation with real data is given in Figure 1.

Since the equations are too long to fit into Table 3, the coefficients $X_{CCH1}$, $X_{CCH2}$, $X_{DCH1}$, $X_{RCH1}$, and $X_{RCH2}$ were used with the full form of these coefficients is given in Appendix 6.1. The initial population in each GP iteration was 455, 225, and 281. The GP was executed for 188 generations for the model of confirmed and deceased cases while the value for the model of recovered cases randomly set to 120. In each generation total of 21, 28, and 47 members were randomly chosen from population to be used in the tournament selection process. The initial tree depth was set in range from 4 or 6 up to 9 to 12 depending on the case as shown in Table 3. All mathematical equations were obtained with very high crossover coefficient value which was between 0.91 and 0.92. Low values of mutation coefficients were selected in each case meaning that the mutation of the candidate solutions did not have as large an influence on the population members in comparison to the crossover operation. Stopping criteria was randomly chosen as 0.469, 0.36, and 0.092. The maximum number of samples were randomly chosen to be 0.81, 0.804, and 0.72. The constant range in GP parameter represents the constant interval ($-3221.2$, $6386.5$). As seen on Figure 1 the mathematical models obtained for confirmed, deceased, and recovered cases have excellent performance since comparison between estimation produced by these mathematical expressions and real data generates $R^2$ score higher than 0.998 as it can be seen on Figure.

### Table 3. Symbolic expression for confirmed, deceased, and recovered cases in China with GP parameters, and $R^2$ score.

| GP parameters Symbolic expression | $R^2$ score |
|----------------------------------|-------------|
| Confirmed $y_{CCH} = \max(X_{CCH1}, X_{CCH2})$ | 0.9991 |
| Deceased $y_{DCH} = \max(X_{DCH1}, -\left[\max(X_0, X_1 - X_2 - X_3^2) + X_0 + X_2^2\right])$ | 0.9992 |
| Recovered $y_{RCH} = 73.7X_{RCH1} - 2.27X_1X_{RCH2}$ | 0.9986 |
Figure 1. Comparison of mathematical models with (a) confirmed cases, (b) deceased cases, (c) recovered cases, and (d) epidemiology curve for China.
The Epidemiology estimation shown in Figure 1(d) shows the high-quality estimation of the GP model on real data.

### Italian model

In this subsection, results for the obtained Italian model are presented. The best equations obtained for each case are given in Table 4 alongside GP parameters that were used to obtain each equation and $R^2$ score of each equation. The comparison of each case with real data and epidemiology trend are shown in Figure 2.

The coefficients $X_{\text{CIT}}$, $X_{\text{DIT}}$, and $X_{\text{RIT}}$ were introduced in order to shorten the notation of mathematical formulas. The full form of these coefficients is given in Appendix 6.2. As seen from Table 4 the best mathematical expressions obtained using GP are selected based on their $R^2$ score. The population of mathematical expressions in each iteration for each case was randomly chosen to be 665, 823, and 606. These numbers are much higher to those used to obtain the Chinese model. The GP trained the models for a maximum number of 170, 171, and 135 generations since the stopping criteria of 0.862, 0.675, and 0.498 were not achieved. In each generations total of 35, 21, and 34 depending on the case population members (mathematical expressions) were competing to become part of the next generation. The crossover coefficient in each case when compared to mutation coefficients was much higher meaning that crossover had much higher influence than mutation coefficient. The stopping criteria in each case was randomly chosen and the values are 0.862, 0.675, and 0.498. This criteria was never met since the GP stopped execution after the maximum number of generations is achieved. The maximum number of samples of training data set in each case that was used was 90%, 97%, and 80%, respectively. The constant range used for the construction of each equation was sizable and, for example, for the model of confirmed cases the constants ranged from $-8825.9$ up to $7587.3$. The parsimony coefficients that are responsible for controlling the program growth from generation to generation were randomly chosen for each case and are equal to 0.065, 0.036, and 0.011.

From Figure 2 it can be seen that mathematical expressions have good approximation on Confirmed, Deceased, and Recovered Cases as well as estimation of epidemiology trend.

### Table 4. Symbolic expression for confirmed, deceased, and recovered cases in Italy with GP parameters, and $R^2$ score.

| Case     | GP parameters | Symbolic expression | $R^2$ score |
|----------|---------------|---------------------|-------------|
| Confirmed| 665, 170, 35, (4, 7), 0.005, 0.005, 0.004, 0.862, 0.9, 0.8825.9, 0.9, 0.065 | $y_{\text{CIT}} = \min \left( X_{\text{CIT}}, \sqrt{-X_0 - X_i - X_0} \right)$ | 0.99975 |
| Deceased | 823, 171, 21, (6, 7), 0.92, 0.0008, 0.01, 0.02, 0.675, 0.97, 0.4010.82, 6373.64, 0.036 | $y_{\text{DIT}} = -X_0^2 + 4X_0 - X_{\text{DIT2}} + X_1^2 + X_{\text{DIT1}}$ | 0.99987 |
| Recovered| 606, 135, 34, (4, 7), 0.91, 0.05, 0.01, 0.005, 0.498, 0.8, 0.5127.2, 6942.3, 0.011 | $y_{\text{RIT}} = \log \left( \log \left( \log \left( \max(3826.99 - X_i, |X_i|) \right) \right) \right) - X_0 + X_1 + X_{\text{RIT1}}$ | 0.99868 |
Figure 2. Comparison of mathematical models with (a) confirmed cases, (b) deceased cases, (c) recovered cases, and (d) epidemiology curve for Italy.
Spain model

Spanish epidemiology trend is somewhat similar to the Italian epidemiology trend. Although the number of confirmed cases each day is shown to be rapidly growing and the number of deceased cases is smaller than in Italian epidemiology model while the number of recovered case is much larger. The equations used for estimation in each case with GP parameters and $R^2$ scores are shown in Table 5. The comparison of mathematical expressions with real data for each case is shown in Figure 3.

In Table 5 the coefficients $X_{CSPA1}, X_{CSPA2}, X_{DSPA1}, X_{DSPA2}, X_{RSPA1}$, and $X_{RSPA2}$ were used to shorten the display of mathematical equations within the table. The full form of these coefficients is given in Appendix 6.3. The numbers of candidate solutions for each presented group are 647, 538, and 623, respectively. The first stopping criteria in GP is the maximum number of generations which is 126, 191, and 198 generations. The other stopping criteria—stopping criteria coefficient, which was in the presented cases set to (0.447, 0.1, and 0.4) was not achieved hence the GP stopped the execution when the first stopping criteria was met that is maximum number of generations. The crossover coefficients are 0.92, 0.92, and 0.92. The next three coefficients in each case represent subtree mutation, hoist mutation and point mutation coefficients and for the confirmed cases shown in Table 5 are 0.04, 0.001, and 0.005, respectively. Again, the crossover operation on the population members is shown to have a greater influence than the mutation coefficients. The maximum number of samples that were randomly selected in training data set between generations, for each model are 91%, 81%, and 95%. The constant range interval that GP used for random selection of constants in order to construct the mathematical expressions range from −4874.4 up to 2609.3 in confirmed case, from −3120.4 up to 2542.8 in deceased cases and from −1459.5 up to 8805.6 in recovered cases. The parsimony coefficients in each case are 0.032, 0.056, and 0.059, respectively. These are very low values which means that the mathematical expressions became very large due to lack of correlation between input and output variables. Note that the size of mathematical expressions cannot be seen from Table 5 but from coefficients given in Appendix.

As seen from Figure 3 this model is similar to the Italian model which means that the number of confirmed and deceased cases is increasing rapidly while the number of recovered cases is slowly increasing. The mathematical expressions from Table 5 that are shown in Figure 3(a)–(c) it can be seen that these mathematical expressions follow the trend confirmed, deceased, and recovered cases really well. In Figure 4(d) all three mathematical expressions were combined as shown.

### Table 5. Symbolic expression for confirmed, deceased, and recovered cases in Spain with GP parameters, and $R^2$ score.

| GP parameters | Symbolic expression | $R^2$ score |
|---------------|---------------------|------------|
| Confirmed     | $y_{CSPA} = \max(X_{CSPA1}\log(\min(2X_1X_2+1139.78),\max(X_0,X_1)))$ | 0.9998 |
| Deceased      | $y_{DSPA} = \max(X_{DSPA1} - X_{DSPA2}\log\left(\frac{X_1}{X_0}\right))$ | 0.9996 |
| Recovered     | $y_{RSPA} = \frac{X_{RSPA1}}{X_{RSPA2}}X_2$ | 0.9993 |
Figure 3. Comparison of mathematical models with (a) confirmed cases, (b) deceased cases, (c) recovered cases, and (d) epidemiology curve for Spain.
in equation (1). The daily number of recovered and deceased cases were subtracted from the daily number of confirmed cases in order to determine the number of active cases. This procedure is done for real data and for the obtained mathematical expressions. From Figure 3 it can be seen that the mathematical expression almost perfectly estimates the real data with only a small deviation.

**USA model**

The USA is one of the last countries in which COVID-19 epidemic began and it is almost impossible to make predictions since the number of confirmed cases and deceased cases is rapidly growing. However, the data collected so far is enough to obtain mathematical expressions for confirmed, deceased, and recovered cases and possibly make the estimation of epidemiology curve using aforementioned expressions. The mathematical expressions for each case with GP parameters that were used to obtain these mathematical expressions and their performance measured in terms of $R^2$ score are shown in Table 6. The comparison of each mathematical expression with real data and the epidemiology trend is shown in Figure 4.

As seen from the Table 6 in order to write mathematical expressions in shorter form the coefficients $X_{CUS1}$, $X_{CUS2}$, $X_{DUS1}$, and $X_{RUS1}$ and the full form of these coefficients is given in Appendix 6.4. The $X_0$ variable in Table 6 represents latitude. The larger sizes of these equations can be attributed to the lack of correlation between input and output data, causing the selected parsimony coefficient to be very small. The randomly chosen parsimony coefficient for each model was equal to 0.05, 0.08, and 0.08, respectively. The number of population members in each case is equal to 141, 181, and 196, respectively. The number of generations which represents one of stopping criteria for each case was set to 199, 104, and 159, respectively. The randomly chosen stopping criteria coefficient was set to the values of 0.864, 0.92, and 0.78, respectively. The value of stopping criteria was never reached so the GP algorithm stopped execution after maximum number of generations was reached. The crossover coefficient (0.77, 0.79, and 0.78) although lower than in previous models is still dominating variational operator when compared to other three mutational operators. The maximum number of samples that were randomly chosen in each generation for each case are equal to 99%, 99%, and 98%. The coefficient range that was used in GP to construct mathematical expressions (population members) are from −7050.5 up to 8861.4 for the model of confirmed cases, from −9534.1 up to 9763.5 for the model of deceased cases, and from −2886.8 up to 7821.8 for the model of recovered cases. As seen from Figure 5, the mathematical expressions for each case perform almost perfectly with smaller deviations when compared to the real data. When the performance of all three mathematical expressions is compared, the mathematical expression for recovered cases shows the largest deviation from the real data. In Figure 5(d) equation (1) is used on real data in order to obtain the number of active cases. The same procedure is applied to the estimation curve with exception that the number of confirmed, deceased, and recovered cases were generated with mathematical expressions presented in Table 6. The estimation generated using these mathematical expressions is almost identical to the real data.

**Global model**

In order to find the global mathematical model that could be utilized for estimation on real data, the data set used for training and testing on confirmed, deceased, and recovered cases must be adjusted. In each data set, the sum of all cases for each day must be calculated in order to determine the total number of confirmed, deceased, and recovered cases. The latitude and longitude as input variable in this model are set to zero since the inclusion of these parameters into the model failed to generate accurate mathematical models. The mathematical models obtained on confirmed, deceased, and recovered cases have one input variable which is number of days since the first data entry. In Table
Figure 4. Comparison of mathematical models with (a) confirmed cases, (b) deceased cases, (c) recovered cases, and (d) epidemiology curve for USA.
the mathematical equations obtained using GP for each case are shown with the GP parameter used to obtain these equations and $R^2$ score that each mathematical equation achieved on the testing data set. The graphical representation of each mathematical equation compared to the real data and estimation epidemiology trend in comparison with real data is shown in Figure 5.

In Table 7 the coefficients $X_{CGLOB1}$, $X_{DGLOB1}$, $X_{RGLOB1}$, $X_{RGLOB2}$, and $X_{RGLOB3}$ were introduced since the mathematical expressions are too large to fit in the table. The $X_2$ represents the number of days from which the data collection for COVID-19 outbreak started. The full form of these coefficients is given in Appendix 6.4. Besides the mathematical expressions the GP parameters which were used to obtain these mathematical expressions were also shown alongside the $R^2$ score. As in previous subsections, all GP parameters were randomly selected. The number of population members for each case equal 374, 439, and 282, respectively. The first stopping criteria is the maximum number of generations which was set to 159 in confirmed case, 180 in deceased case, and 156 in recovered case. The second stopping criteria was set for each case to 0.63, 0.8, and 0.8, respectively. Since the second stopping criteria represents the minimum $\text{MAE}$ which is the fitness value of the best population member in the given generation and was never achieved the GP algorithm stopped the training process when the maximum number of generations was achieved. In each generation the total of 37, 21, and 20 population members were competing against each other to become the members of the next generation. As in previous cases the half-and-half method was utilized to construct the initial population with tree depth in range from 4 to 12 in confirmed case, 4 to 10 in deceased, and 3 to 8 in recovered case receptively. The crossover coefficient in each case with values of 0.73, 0.72, and 0.73 is still the dominating EC operator when compared to the values of the three mutation coefficients. The maximum number of samples that were randomly picked from training data set in every generation for each case was 99%, 96%, and 92%, respectively. The coefficient range which was used in each case by the GP algorithm to construct population members or to perform crossover or mutation was set from $-6337.8$ up to $5006.5$ for the model of confirmed cases, from $-4511.08$ up to $2513.46$ for the model of deceased cases, and from $-7246.9$ up to $6715.4$ for the model of recovered cases. The parsimony coefficients are set to very low values which means that the population members could grow in length and depth from generation to generation. The results of the low parsimony coefficients are very large mathematical expressions that are able to correlate input with the output variable.

In Figure 5(a)–(c) the performance of mathematical expressions for confirmed, deceased, and recovered case are compared to real data. The comparison showed that mathematical expressions provide good approximations to the real data for confirmed and deceased models, with the mathematical equation for recovered case showing some deviations from the real data. In Figure 5 using equation (1) the epidemiology trend estimation is made. The real data from confirmed, deceased, and recovered cases models was used to determine the number of active cases using equation (1). Then the data generated by three equations from Table 7 was used in equation (1) in order to obtain the number of infected cases. The comparison of these two curves in Figure 5(d) shows that the epidemiology curve based on these three equations shows a good approximation of real epidemiology data.
Figure 5. Comparison of mathematical models with (a) confirmed cases, (b) deceased cases, (c) recovered cases, and (d) epidemiology curve on global scale.
Discussion

In Table 1 the possible ranges of hyperparameters used in GP algorithm are shown. The crossover probabilities for all models tended to higher values, with all mutation probabilities tending to lower values. Because of this, crossover probability had a bigger influence on population of symbolic expressions in each generation than the mutation coefficient probabilities. Stopping criteria coefficient was never achieved, because of a very low upper bound of values from which it could be selected. Due to this, the secondary stopping criteria, number of generations, was used to stop the training process in all models. The number of generations shows tendency towards the upper bound (200) across models. The maximum number of samples used in training shows the tendency to the upper bound of possible values. The resulting range of parsimony coefficients which were used to generate the best symbolic expressions is very low. This happened due to weak correlation between input and output variables, in order to enable growth of symbolic expressions in terms of achieving lower MAE value. The other hyperparameters show equal distribution amongst possible values, not showing a tendency to either bound.

Today a vast number of AI algorithms exist which can be used to solve specific problems. The most popular AI algorithms are ANNs which are trained in a similar manner to GP in order to solve specific problem. The result of training the ANN to solve a specific problem is the architecture which is capable of solving the problem. The transformation of ANN architecture into mathematical expression is almost impossible due to the large number of neurons and their interconnections. On the other hand the benefit of using GP algorithm is it will, after training, produce the mathematical expression that correlates inputs and the desired output.

In order to obtain epidemiology trend equation for the Chinese, Italian, Spanish and USA models GP was utilized to obtain symbolic expressions for the number of confirmed, deceased, and recovered cases. For each symbolic expression obtained using GP the dataset consisted of latitude, longitude, and a number of confirmed/deceased/recovered cases for each day since the dataset start. Due to a small dataset at the moment of the research the latitude and longitude were fixed for each specific location in each model. The Chinese model is the most unique model when compared to other modes since the outbreak of COVID-19 started earlier than in other models and the number of confirmed and deceased cases is stagnating while the number of recovered cases is slowly growing. The latitude and longitude for the Chinese model was fixed at Hubei province (latitude: 31, longitude: 112) due to a small numbers of other locations for reported confirmed/deceased/recovered cases. The best symbolic expressions for confirmed, deceased, and recovered cases obtained using GP generated $R^2$ score values of 0.999, 0.999, and 0.998, respectively. These results showed that these symbolic expressions estimate the confirmed, deceased, and recovered cases with high accuracy when compared to real data. Combining the best symbolic expressions in terms of $R^2$ score for confirmed/recovered/deceased cases generated epidemiology trend equation which follows the real epidemiology curve with high accuracy which can be seen from Figure 1. From that figure, two interesting features can be noticed: after 20 days since the
start date number of confirmed cases rapidly grew while the number of deceased and recovered cases had slowly increased, after 25 days since the start date the epidemiology curve achieved maximum value and started to decrease. The rapid increase in number of confirmed cases can be attributed to the fact that the strict measures of social distancing and hygiene were not fully implemented. The decrease in epidemiology curve can be attributed to the fact that the number of confirmed cases started to stagnate while the number of deceased and recovered cases are slowly growing. The epidemiology curve showed that the number of infected cases dropped down below 1000 after 70 days from the start date. The decrease in number of infected cases is a result of strict quarantine measures followed by obligatory protection equipment as well as social distancing and extremely high hygiene and sanitation standards.

For the Italian model, the symbolic expressions for confirmed/deceased/recovered cases were obtained using the same procedure as described in case of the Chinese model. Since at the time the symbolic expressions were obtained there were small numbers of locations in which COVID-19 was confirmed and they were mostly concentrated at hospitals the location in terms of latitude and longitude values was fixed at the city of Milan (latitude: 45.27, longitude: 9.11). When compared to the Chinese model the Italian model had exponential growth of the number of confirmed/deceased/recovered cases. This growing trend can be attributed to the fact that extreme quarantine measures, social distancing and high sanitation standards were not implemented as they should have been. The best symbolic expressions for confirmed/deceased/recovered cases obtained using GP generated $R^2$ score values of 0.9991, 0.999, and 0.998, respectively. As seen in Figure 2 all these symbolic expressions are following the real trend of confirmed, deceased, and recovered cases with high accuracy. Using these equations the epidemiology equation was generated and when compared to the real trend it can be seen that this equation estimates the real trend with high accuracy. At the time of conducted investigation the Italian epidemiology trend was exponentially growing without any indication of decreasing. In order to decrease number of confirmed/deceased cases and increase the number of recovered cases, Italy will have to introduce strict measures of quarantine, social distancing, and high hygiene and sanitation standards.

The epidemiology trend in Spanish model shows the same behavior as the Italian model, with exception of the number of confirmed cases, the growth of which was slowly decreasing. The procedure for obtaining the symbolic expressions for confirmed, deceased, and recovered cases is the same as in previous models. Due to the small number of reported locations the latitude and longitude values were fixed to the city of Madrid (latitude: 40.38, longitude: −3.71). The best symbolic expressions using GP for confirmed, deceased, and recovered cases achieved $R^2$ values of 0.999, 0.999, and 0.999, respectively. These values showed that these symbolic expressions are estimating the number of confirmed, deceased, and recovered cases with high accuracy. The epidemiology equation was formulated using the best symbolic expressions for confirmed, deceased, and recovered cases. From Figure 3 it can be seen that the epidemiology equation of Spanish model is estimating the number of infected cases with high accuracy when compared to the real data. From the epidemiology curve it can be noticed that the number of infected cases is slowly increasing. This trend can be attributed to the fact that after 70 days since the start date was recorded the number of confirmed and deceased cases showed a slightly slower increasing trend while the number of recovered cases was constantly growing. As mentioned previously, the Spanish model is similar to the Italian model with the exception of the number of infected cases which is slowly increasing after 70 days since the initial data was recorded in the data set. This means that the extreme quarantine measures as well as social distancing, high hygiene and sanitation standards were implemented in full form.

When compared to the other countries in the USA the outbreak has recently started. However, due to high number of violations of extreme quarantine measures as well as social distancing, hygiene, and sanitation measures the number of confirmed and deceased cases exponentially increased while the number of recovered cases showed only a slow increase. Again, the procedure for obtaining symbolic expressions for confirmed, deceased, and recovered cases using GP is the same as in previous models.
and due to a small number of reported locations the latitude and longitude values were fixed at the city of New York (latitude: 40.716, longitude: −74). The best symbolic expressions obtained for confirmed, deceased, and recovered cases were obtained using GP that achieved $R^2$ values of 0.999, 0.998, and 0.995, respectively. All three equations are estimating the number of aforementioned cases with high accuracy which can be seen in Figure 4. These three equations were used to formulate epidemiology trend equation and the results showed that this equation is estimating the number of infected cases with high accuracy when compared to the real epidemiology curve. From the real data shown in Figure 4 it can be seen that the number of infected cases is still rapidly growing which can be attributed to the fact that general population is not following strict measures from World Health Organization (WHO) or Center for Disease Control (CDC). In order to stop this trend, the general population should follow extreme measures of quarantine as well as social distancing and high hygiene and sanitation standards. If not the consequences could be catastrophic.

The epidemiology trend equation on global scale was obtained following the same procedure as in the case of Chinese, Italian, Spanish, and USA model. However, the latitude and longitude were omitted due to small number of outbreak locations, at the time the analysis was conducted. So the only input variable was the number of confirmed/deceased/recovered cases for each day since the start date. The best symbolic expressions obtained for three cases using GP achieved $R^2$ score values of 0.999, 0.999, and 0.983, respectively. These values indicated that the obtained symbolic expressions are estimating the numbers of confirmed, deceased, and recovered cases with very high accuracy when compared to the real data. Using these symbolic expressions the epidemiology equation was formulated for global model and from Figure 5 it can be seen that the estimated epidemiology trend is following the real epidemiology trend with high accuracy. From estimated and real epidemiology curve two interesting features can be noticed. After 20 days since the data collection started on the global scale the stagnation of the number of infected cases can be noticed. After 50 days since the start date the number of infected cases started to rapidly increase which can be attributed to the fact that outbreak had started in Italy, Spain, USA and other countries. The epidemiology trend should start to decrease if quarantine measures, social distancing as well as high hygiene and sanitation standards defined by WHO are followed globally.

**Conclusion**

In this paper, the GP algorithm was utilized in order to obtain mathematical expressions for confirmed, deceased, and recovered cases for China, Italy, Spain, USA, and for the entire globe. In each model, the best three equations for confirmed, deceased, and recovered cases are combined together to obtain mathematical expression which could estimate the epidemiology trend. From presented results the following conclusions can be drawn.

- GP algorithm can be utilized to obtain mathematical expressions for estimation of confirmed, deceased, and recovered cases of a specific country with high accuracy.
- For each country model the obtained mathematical expressions could be combined together in order to estimate the epidemiology trend with high accuracy.
- The obtained symbolic expressions for confirmed, recovered and deceased cases, as well as mathematical equations, for estimation of epidemiology trend in Hubei province (China model) is estimating the number of confirmed, deceased, and recovered cases as well as epidemiology curve with high accuracy. From China epidemiology trend it can be seen that the number of infected cases is decreasing which means that the extreme epidemiological measures defined by WHO were implemented and helped with lowering the number of active COVID-19 cases.
• The symbolic expressions obtained using GP algorithm for confirmed, deceased, and recovered cases as well as epidemiology equation in Italian model estimate the number of confirmed, deceased, recovered, and infected cases with high accuracy. The epidemiology curve of Italy is still increasing which means that from the day the outbreak has started the epidemiological measures defined by WHO such as quarantine, social distancing and hygiene are violated by general population.

• The Spanish epidemiology trend is similar to Italian with the exception that after 70 days since the start date the epidemiology curve is slow increasing trend. This change in epidemiology trend can be attributed to the fact that the general population is following the epidemiology measures defined WHO. The symbolic expressions obtained with GP algorithm as well as mathematical expression for epidemiology curve are estimating whit high accuracy when compared to the real data.

• The symbolic expressions obtained using GP algorithm as well as mathematical expression for epidemiology trend for the USA are estimating the number of confirmed, deceased, recovered cases as well as the real epidemiology curve with high accuracy. The USA epidemiology trend is one of the most concerning trends when compared to the other epidemiology models. Since the outbreak has started the number of infected cases is exponentially increasing which means that general population is violating the epidemiology measures defined by WHO.

• The symbolic expressions using GP algorithm as well as the mathematical expressions for epidemiology estimation on a global scale showed that these equations could estimate then number of confirmed, deceased, and recovered cases with high accuracy.

The presented research shows that by using the GP algorithm the accurate symbolic expressions can be obtained for estimation of the number of confirmed, deceased, and recovered cases. This points towards the possible use of GP in this and similar future epidemics. This investigation also showed that the symbolic expressions for confirmed, deceased, and recovered cases could be used together in order to formulate the mathematical expression for epidemiology trend estimation which could estimate the epidemiology trend with higher accuracy. Authors hope that the hyperparameters of the obtained models along with presented methodology can lay further groundwork for re-fitting models using GP or similar methods with newly gathered data in the future, in both the ongoing pandemic, and similar future challenges. The findings that the models can be generated with the relatively low amount of data used at the time this research was performed points towards the capability of using GP algorithm in beginning stages of epidemics to generate initial spread models.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been (partly) supported by the CEEPUS network CIII-HR-0108, European Regional Development Fund under the grant KK.01.1.1.01.0009(DATACROSS), project CEKOM under the grant KK.01.2.2.03.0004, CEI project “COVIDAi” (305.6019-20) and University of Rijeka scientific grant uniritehnic-18-275-1447.

ORCID iD
Sandi Baressi Šegota https://orcid.org/0000-0002-3015-1024
References

1. Liu Y, Gayle AA, Wilder-Smith A., et al. The reproductive number of covid-19 is higher compared to sars coronavirus. J Travel Med 2020; 27(2): taaa021.
2. Al-Gheethi A, Noman E, Al-Maqtari QA, et al. Novel coronavirus (2019-ncov) outbreak; a systematic review for published papers, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3537085 (2020, accessed 2 September 2020).
3. Zhou P, Yang X-L, Wang X-G., et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. BioRxiv 2020.
4. Ming W-K, Huang J and Zhang CJ. Breaking down of healthcare system: mathematical modelling for controlling the novel coronavirus (2019-ncov) outbreak in Wuhan, China. bioRxiv 2020.
5. Paterlini M. On the front lines of coronavirus: the italian response to covid-19. BMJ 2020; 368: m1065.
6. Tanne JH, Hayasaki E, Zastrow M, et al. Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. BMJ 2020; 368: m1090.
7. Ma J. Estimating epidemic exponential growth rate and basic reproduction number. Infect Dis Model 2020; 5: 129–141.
8. Verelst F, Kuylen E and Beutels P. Indications for healthcare surge capacity in European countries facing an exponential increase in coronavirus disease (COVID-19) cases. Euro Surveill 2020; 25(13): e2000323.
9. Spresco A, Eriksson O, Dahlström Ö, et al. Evaluation of nowcasting for detecting and predicting local influenza epidemics, Sweden, 2009–2014. Emerg Infect Dis 2018; 24(10): 1868.
10. Vynnycky E and White R. An introduction to infectious disease modelling. Oxford: Oxford University Press, 2010.
11. González-Crespo R, Herrera-Viedma E and Dey N. Finding an accurate early forecasting model from small dataset: A case of 2019-nCoV novel coronavirus outbreak. Int J Interact Multimed Artif Intell 2020; 6(1): 132–140.
12. Lorenzin I, Anelić N, Španjol J, et al. Using multi-layer perceptron with laplacian edge detector for bladder cancer diagnosis. Artif Intell Med 2020; 102: 101746.
13. Richardson P, Griffin I, Tucker C, et al. Baricitinib as potential treatment for 2019-ncov acute respiratory disease. The Lancet 2020; 395(10223): e30–e31.
25. Jit M and Brisson M. Modelling the epidemiology of infectious diseases for decision analysis. *Pharmacoconomics* 2011; 29(5): 371–386.

26. De Angelis D, Presanis AM, Birrell PJ, et al. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* 2015; 10: 83–87.

27. Dietz K and Schenzle D. Mathematical models for infectious disease statistics. In: *A celebration of statistics*. Cham: Springer, 1985, pp.167–204.

28. Grassly NC and Fraser C. Mathematical models of infectious disease transmission. *Nat Rev Microbiol* 2008; 6(6): 477–487.

29. Höhle M, Jørgensen E and O’Neill PD. Inference in disease transmission experiments by using stochastic epidemic models. *J R Stat Soc Ser C Appl Stat* 2005; 54(2): 349–366.

30. O’Neill PD. Introduction and snapshot review: relating infectious disease transmission models to data. *Stat Med* 2010; 29(20): 2069–2077.

31. Tsui K-L, Wong ZS-Y, Goldsman D, et al. Tracking infectious disease spread for global pandemic containment. *IEEE Intell Syst* 2013; 28(6): 60–64.

32. Wong ZS, Zhou J and Zhang Q. Artificial intelligence for infectious disease big data analytics. *Infect Dis Health* 2019; 24(1): 44–48.

33. Schwartz R, Dodge J, Smith NA, et al. Green ai. *arXiv preprint arXiv:1907.10597* 2019.

34. Kay J. Ai and education: grand challenges. *IEEE Intell Syst* 2012; 27(5): 66–69.

35. Chan KS and Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ* 2019; 5(1): e13930.

36. Car Z, Šegota SB, Andelić N, et al. Modeling the spread of COVID-19 infection using a multilayer perceptron. *Comput Math Methods Med* 2020: e5714714. https://doi.org/10.1155/2020/5714714

37. Santosh K. Ai-driven tools for coronavirus outbreak: need of active learning and cross-population test models on multidimensional multimodal data. *J Med Syst* 2020; 44(5): 1–5.

38. Costa FF. Big data in biomedicine. *Drug Discov today* 2014; 19(4): 433–440.

39. Yang Y, Zheng X, Guo W, et al. Privacy-preserving smart iot-based healthcare big data storage and self-adaptive access control system. *Inform Sci* 2019; 479: 567–592.

40. Rajovic N, Rico A, Puzovic N, et al. Tibidabo: making the case for an arm-based hpc system. *Future Gener Comput Syst* 2014; 36: 322–334.

41. Koza JR and Koza JR. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. Cambridge: MIT Press, 1992.

42. Uy NQ, Hoai NX, O’Neill M, et al. Semantically-based crossover in genetic programming: application to real-valued symbolic regression. *Genet Program Evolvable Mach* 2011; 12(2): 91–119.

43. Walker M. *Introduction to genetic programming*. Tech. Np: University of Montana, 2001.

44. Cormen TH, Leiserson CE, Rivest RL, et al. *Introduction to algorithms*. Cambridge: MIT Press, 2009.

45. Langdon WB and Banzhaf W. Faster genetic programming gpquick via multicore and advanced vector extensions. *arXiv preprint arXiv:1902.09215* 2019.

46. Koza JR. *Genetic programming II*, volume 17. Cambridge: MIT Press, 1994.

47. Eiben AE and Smith JE. *Introduction to evolutionary computing*, volume 53. Cham: Springer, 2003.

48. Koza JR, Andre D, Keane MA, et al. *Genetic programming III: Darwinian invention and problem solving*, volume 3. Burlington: Morgan Kaufmann, 1999.

49. Jha SK and Eyong EM. An energy optimization in wireless sensor networks by using genetic algorithm. *Telecommun Syst* 2018; 67(1): 113–121.

50. Kim H, Kim S-H, Jeon M, et al. A study on path optimization method of an unmanned surface vehicle under environmental loads using genetic algorithm. *Ocean Eng* 2017; 142: 616–624.

51. Goodfellow I, Bengio Y and Courville A. *Deep learning*. Cambridge: MIT Press, 2016.

52. Lorencin I, Anelić N, Mržljak V, et al. Genetic algorithm approach to design of multi-layer perceptron for combined cycle power plant electric power output estimation. *Energies* 2019; 12(22): 4352.

53. Huang J, Cen Y, Xie N, et al. Inverse calculation of demolition robot based on gravitational search algorithm and differential evolution neural network. *Int J Adv Robot Syst* 2020; 17(3): 1729881420925298.

54. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Berlin: Springer Science & Business Media, 2009.
In this section all the coefficients for China, Italy, Spain, USA and global model that were introduced in order to simplify the best symbolic expressions for confirmed, deceased, and recovered
model are provided. In all the equations variable $X_0$ represents latitude, $X_1$ represents longitude and $X_2$ represents the number of days since the outbreak began.

China model

$$X_{China} = (4864.15 \max(-2.65181\sqrt{X_2} - \sqrt{\frac{X_2}{X_2 - X_0}} - X_0 - X_1, -4759.76) / \\
((X_2^{3/2} \sqrt{\frac{X_2}{(X_2 - X_0)^{3/2}}} - \sqrt{\frac{X_2}{(X_2 - X_0)^{3/2}}} - X_0 - X_1, -4759.76) / \\
(|\min(\sqrt{X_2} - X_0, \log(\sqrt{X_2} - X_0) - \sqrt{X_2} - X_0)|)^{1/6}) / \\
((X_2^{3/2} - X_0)^{3/2} + X_2)) - \sqrt{X_2 - X_0}) / (|X_0|(|\sqrt{X_0}|)\sqrt{X_2 - X_0}) \\
((X_2^{3/2}(\log(-X_1 + |2425.44X_0(X_2 - X_0)|\log(\sqrt{X_2} + X_2))) / \\
(\min(X_2 - X_0, 0.00016X_0)) / \\
\left(\left(X_2 - X_0\right)^{3/4} + [\frac{X_0}{X_1X_2}]\right)^{1/2} + 7.03\sqrt{\max\left(\sqrt{X_2} - \sqrt{X_2 - X_0}, \log(7.03X_0)\right)} \\
|+4309.08)^{1/2}/((X_2 - X_0)^{3/2})^{1/2}(|\min(\log(|3.38X_0|)|\sqrt{X_0}|) \\
\sqrt{\frac{X_2^{3/2}\log(-X_1 + 7.03\sqrt{X_2} + 4309.08)}{(X_2 - X_0)^{3/2}}} \left| X_2 - X_0 \right| \\
\log(X_0 - 1480.91) \max(6016.38, X_2)X_2^3)) / ((X_2 - X_0)^{7/2}(\sqrt{X_2}X_2 \\
+ \sqrt{X_2} - \log\left(\sqrt{X_2} - X_0\right)^{1/2})^2 + 7.03\sqrt{\frac{X_2}{X_2 - X_0}} + |\sqrt{X_0}|, \max\left(\frac{7.03\sqrt{X_2}}{\sqrt{X_2} - X_0}, \\
\frac{\sqrt{X_2} - X_0 - X_0 + X_2}{\left(\sqrt{X_2} - X_0\right)^{1/2}} - (\min(\sqrt{X_2}, \log(-X_2))(\sqrt{X_2} \\
- \log(\sqrt{X_2} - X_0) + X_2 + \sqrt{X_2} + X_0 - 6285.55)) + 7.03\sqrt{X_2}\log(\min(-X_0 \\
- X_1 + X_2 + 4309.08, \log(X_2)(\min(\sqrt{X_2} - X_0, \log(X_2 - X_0) \\
- \sqrt{X_2 - X_0})^{1/2} + X_0 + (4309.08 - X_1)(4845.67 - 2425.44|X_0|))))

(6.1)
\[ X_{\text{ccm2}} = 0.142205(4309.08 - X_1) \left( \frac{2.65181 \sqrt{X_2}}{\max(6016.38, \log(X_2 - X_0))} + X_2 \right) \] (6.2)

\[ X_{\text{dch1}} = -\max\left( \frac{X_1 \sqrt{X_2}}{\max(-2245.39 X_2, X_2) - X_2 |X_0|}, \min(1597.94, -X_0) \right) \]
\[ \sqrt{\min\left( \max(1444.56, X_1), \sqrt{\min(-1597.94, X_0)} \right)} \]
\[ -\max\left( X_2, \frac{X_2^2}{\log(X_0 - X_1 + X_2)} \right) - |X_2^2||) \]
\[ \left( \min\left( X_2 |X_0|, \frac{\log(X_0)}{X_2} \right) \right) / \left( -\max(X_2, \frac{X_2^2}{\sqrt{\min(-X_0, |X_0|)}}) \right), \]
\[ \min(X_2 |X_0|, \sqrt{|X_0|} - |X_1 - X_2|^2), (\max(272.12 X_1 (-X_0 - X_2 + 71.9),
\[ |X_0|^{1/2}) - X_0 - 29.9 X_1) \] (6.3)
\[ X_{Rh1} = \left( \min(73.7, 73.7 \max(\min(2006.9, X_2), \min(2805.35, X_2))) \right) \]

\[ - \min(2805.35, |X_2|) + X_0 + X_1, \right) \left( \min(-X_2, \min(2006.9, -X_2, X_2)) \right) \]

\[ \min(2006.9, X_2, \max(2006.9, X_2)) - \min(2805.35, |X_2| + |X_0 + X_1|) / \left( |X_0| + \sqrt{X_1 - \max(19.0614, |X_0 + X_1| - |X_2| - X_2)} \right) \max \]

\[ \left( \min(X_1 + X_2, |X_0|) - \min(2805.35, |X_2| + |X_0 + X_1|), \right) \max(\left( \left( \min(-X_2, |X_1 + X_0|) \right) / \left( \sqrt{X_2} \right) \right) \left( \sqrt{X_1} \right) \left( \min(73.7, \sqrt{X_1 - 73.7, X_1 + X_2, |X_0|}) \right) \left( \sqrt{X_2} \right) \right) + \left( \min(2006.9, X_2, \max(|\log(|X_0 + X_1|)| + (X_1 - (-X_2 + |X_0 + X_1| - |X_2|)) \right) \max \left( \left(-73.7, |\min(2006.9, -X_2, X_2)| - \min(2805.35, |X_0 + X_1|) \right)^{1/2}, \right) \min(X_1, X_2, |X_0|) \) \left( -\min(2805.35, |X_2| + |X_0 + X_1|), \right) \min(2006.9, -X_2, \max((X_1 - (|X_0 + X_1| - |X_2| - X_2) \right) \max(-73.7, \max(-73.7, |X_2|) \left( \min(X_0 + X_1, X_2) + X_0 + X_1 - |X_2| \right))^{1/2}

\[ + |\log(|X_0 + X_1|)|, \min(X_1, X_2, |X_0|)\) - 724.708 \}^{1/2}

\[ + \sqrt{X_0 + X_1}, \min(X_1 + X_2, |X_0|) \right) / \left( |X_0| + \sqrt{-73.706 \min(2006.9, X_2, 2805.35 X_2) + X_1 + 73.706} \right) \]

\[ X_{Rh2} = \sqrt{\frac{1}{X_1} \min(73.706, X_1 + X_2, \sqrt{X_1 - X_1 X_2, |X_0|})} \left( \sqrt{\frac{1}{X_1} \sqrt{X_2}} \right) \]

Italian model
\[ X_{CIT1} = -((\max(3402.7, X_0) + 0.999289 X_0 + X_1 - X_2) \\
(\frac{X_0}{X_1} - (X_2 - 7085.74)(\min(5454.61, X_2) - X_0 - X_1)) \\
/ (\left| X_2 + \max(3402.7, X_0) + 902.164 \right| - (4304.38 - \sqrt[3]{X_1}((X_0 + X_1) \\
\max(-8618.94, X_2, (-\sqrt[3]{X_1}((-X_0 - X_1) \max \left( \min(3028.61, X_0) + X_0, (-X_0 - X_1) \right) \\
\left( \left( \min(5454.61, X_2) - X_0 \right) \right) / \left( \max(3402.7, X_0) - \frac{X_0}{X_1} \right) \right)

- (\min(5454.61, X_2)) - \frac{4304.38}{X_1})) - (2X_0(0.0007 \left| X_0 \right| - (X_2 - 7086)) \\
\min((-X_0 - X_1) (\left| X_2 \right| - (X_2 - 7085.74)(\min(5454.61, X_2) - X_0 - X_1)) \\
\max(X_0, \min(\min(5454.61, X_2) - X_0 - X_1)) \max(3402.7, X_0) - (\left| X_2 - X_1 \right|)((X_2 - 7085.74) \\
\max(3402.7, X_0) - \frac{4304.38}{X_1}) / \left( \min(5454.61, X_0) - \frac{4304.38}{X_1} \right), \right)

\left(\frac{\max(3402.7, X_0) + \frac{X_0}{X_1}(2X_0 + 2X_1) - 0.0007X_0 - X_2)(\max(3402, X_0)}{X_0 - 7085})(\min(5454, X_2) - X_0 - X_1)\right) \frac{1}{3} \right) / \left( \sqrt[3]{-X_2 + \max(3402.7, X_0) + 902.164} \right) \frac{1}{3} \right) / \left( \min(3402.7, X_0) - \frac{4304.38}{X_1} \right) \left( \max(3402.7, X_0) \right) / \left( \left| X_0 - X_1 \right| \right)

\left( X_0 - 7085.74)(\left| X_0 - X_1 + \min(5454.61, X_2) \right| - \frac{X_0}{X_1} \right) \right)

\left( (2X_0(\max(-8618.94, X_2, (-\sqrt[3]{X_1}((-X_0 - X_1) - \frac{X_0}{X_1} + \max(3402.7, X_0) \frac{1}{3} \right))^2) \\
+ X_0) + \max(-8618.94, X_2))(\min(3402.7, X_0 + X_1) \right) - \left( X_2 - 7085.74)(\min(5454.61, X_2) - X_0 - X_1) \right) / \left( \left| X_0 - X_1 + X_2 + 902.164 \right| \right) - (\left| X_2 - X_1 \right|)((X_2 - 7085.74) \\
\max(3402.7, X_0 - X_0 - X_1)) + \right) \\
\frac{X_0}{X_0 + X_1} + X_0 + X_1 - X_2) / \left( \max(3402.7, X_0) - \frac{X_0}{X_1} \right) + X_2 + \left( \max(3402.7, X_0) + 902.164 \right) / X_1 \right) \frac{1}{3} \} / X_1} \]
\[ X_{DIT1} = \max(-\max(X_0, X_2) - X_2, \max(X_2, \max(626.228, -2X_0, \left| X_0 + \sqrt{X_1}\right| X_2 + X_2^2 - 2865.61, \max(X_0X_2, -X_0 (626.228, -X_1X_2))) + X_0) - \left| X_0 \right|^2 \frac{(X_0 - X_2)(X_1X_2 + X_2 - 626.228)}{X_1}) + 3X_0 + X_2^2) - 2865.61, -\min(\frac{(X_0 - X_2)(X_0 + \sqrt{X_1} - X_2 - 2865.61)}{X_1}, X_2, X_0 (626.228, -X_1X_2) + 626.228) \]

\[ X_{DIT2} = \max(X_2, \max(\left| X_2^2 + |X_2|^2 \right|, X_0, X_2) \log(X_1)(626.228 - X_1X_2), ((X_0 - X_2)(X_0 - \max(X_0, \max(-X_0, X_0^2, (X_0 - X_2) \log(X_1)(626.228, -X_1X_2)) - 2865.61) - |X_0|^2 - 2X_0X_2 + X_0 - \frac{2865.61}{X_1})))) / \]

\[ \left( \max(X_1, (X_0 - X_2) \log(X_1)) \right) - X_2^2 - \sqrt{X_1} - 2865.61) - |X_0|^2 - 2X_0X_2 + X_0 - \frac{2865.61}{X_1} \]

\[ X_{RIT1} = \max(\max(X_0 + X_1, \min(X_2 \max((X_0 - X_2) \log(\log(X_0(X_2 - X_0) \max(\sqrt{\log(X_0 - X_2)}, (\log(X_0X_1) + X_0 - X_2) \log(\min(-2X_0 + X_1 + X_2 + 601.16, ((X_0 - X_2 + 7.8)(X_2 - X_0) (-X_2 \left(\log(\sqrt{X_2 - X_0}) + \sqrt{X_1} + X_1 + X_2 + 601.16))^{\frac{1}{2}})))) + X_0 - X_2), (X_0 \log(\min(X_0 + X_1 - 6663.68, (X_2 - X_0) (\min(X_0, X_1) - X_0X_1 + X_1)) + 137.46)^{\frac{1}{2}}), \max((X_2 - X_0) \max(\sqrt{X_1}, X_1 - X_2), (X_2 - X_0) \max(X_2, ((-X_0 + X_1 + 601)(X_2 - X_0) (\log(\sqrt{X_0 + \log(X_0 - 601)} - X_2) + X_0 - X_2))^{\frac{1}{2}}) + \log(\log(\left|X_2\right|))) + \log(\log(\log(X_2 - X_0)))) \)

\[ \left( \log(X_0(2476.71 - X_1)) + \log(X_0X_1) \right) + \log(\sqrt{X_1}), \min(\max(6665.81(X_2 - X_0), |X_1|) + \log(\log(6663.68 - X_0))), X_2 \max(|X_1|, \log(X_2 - X_0))) \]
Spain model

\[
X_{CSPA} = \left| \left( \sqrt{\min \left( -X_0X_2, \max \left( -4354.34, X_0 \right) \right)} \max \left( X_0, \frac{-1255.34}{X_1} \right) \right) 
- \log(X_2) \log(\min(\sqrt{X_1},(28.4336 \max (X_1, X_2)) / \left( (\min(-X_0X_2, \frac{28.44}{\min(X_0, X_1)}) \right)^{\frac{1}{2}} \min(\max(\sqrt{X_0X_2}, \log(2 \log(X_0))))) \right. 
\left. \min\left( \frac{2047.69 \log(X_0)}{X_0}, 0.035 \max(X_1, X_2) \right) + \min(\max(X_1, \log(X_0) + X_1)), \max(2X_2, \log(X_2))) \right)\right| \max(\log(\min(X_0X_2, \sqrt{X_1}))) \right)
\log(\min(2X_1(X_2 + 1139.78), \log(\min(2X_1^2 \log(X_0))) \log(\min(X_1, X_2)), \log(X_0) \min(X_0X_2, 0.035 \max (X_1, X_2)) \log(\log(X_0)))) / (\min(\max(X_1, X_2, 0.035 \max (X_1, X_2)), \log(X_0)) \min(\max(-4354.34, X_0), \min(-X_1 - 1439.46), \max(X_1, X_2), \log(X_0)) \min(3.34, \min(\log(X_0) \log(X_1), 0.035 \sqrt{\min(-X_0X_2, \frac{4.50(X_2 + 1139.78)}{X_0}}) \right) \sqrt{\min(-X_0X_2, \frac{4.51 \log(X_0)}{X_0})} \left) \right)^{\frac{1}{2}}, \min(0.035 \max(X_1, X_2), \log(X_0)) \min(X_0X_2, 0.035 \max(X_1, X_2)) + \min(-\frac{\log(X_1)}{-X_1 - 1439.46}, \max(\max(X_1, X_2), \log(X_0) \min(-4354.34, X_0), \min(0.031 \max(-4354.34, X_0), \max(X_1, \log(\log(X_0)))) / (\min(X_1, X_2)) \log(X_0) \sqrt{\min(28.4336, -X_0X_2) \min(-X_1 - 1439.46, \log(X_0)) \min(X_0X_2, 0.035 \max(X_1, X_2)) \min(2X_1(X_2 + 1139.78), \log(\log(X_0)))) / (\min(X_0 + X_1 \frac{454.145}{X_0} \frac{+ \min\left( \frac{2047.69 \log(X_0)}{X_0}, 0.035 \max(28.4336, X_2) \right) + \min\left( \frac{-4354.34, X_0}{-X_1 - 1439.46}, \max(X_2 + 1139.78, \log(\min(\log(X_0)) \log(X_2), \log(\max(X_0, X_2))) / (\min(X_0 + X_1 - 454.145), \max\left( \frac{2047.69}{X_0}, X_2 + 2278.35 \right), \max(X_0, X_2)) \right)
\[ X_{CSPA}^{2} = \sqrt{\min (X_{0}, -X_{0}X_{2}) \left( \min \left( \frac{2047.69 \log(X_{0})}{X_{0}}, 0.035 \right) \right) - \min \left( -X_{0}X_{2}, \frac{28.4}{\min(X_{0}, X_{1})} \right) \max \left( \frac{\sqrt[4]{X_{0}X_{2}}, \log \left( \frac{\log(X_{0})}{X_{0}} + 4703.63 \right)}{X_{0}} \right), \] 
\[ \min \left( \frac{2047.69 \log(X_{0})}{X_{0}}, 0.035 \max (X_{1}, X_{2}) \right) \max \left( 2X_{2}, \log(X_{2}) \right) \min \left( \frac{\max \left( -4354.34, X_{0} \right)}{-X_{1} - 1439.46}, \max \left( \frac{2047.69}{X_{0}}, X_{1} \right) \right) \right) \] 
\[ \min \left( \frac{\max \left( -4354.34, X_{0} \right)}{-X_{1} - 1439.46}, \max (X_{1}, \log(X_{1} + \log(X_{0}))) \right) \sqrt{\min \left( X_{0} - X_{0}X_{2} \right)} \] 
\[ \left( \min \left( \frac{2047.69}{X_{0}}, X_{1} \right), \max (X_{1}, \log \left( \frac{\sqrt[4]{X_{0}X_{2}}, \log \left( \frac{\log(X_{0})}{X_{0}} + 4703.63 \right)}{X_{0}} \right)) \right) \] 
\[ \max \left( 2X_{1}, X_{2} + 1139.78, \log \left( \log(X_{0}) \right), \max \left( \frac{\sqrt[4]{X_{0}X_{2}}, \log \left( \frac{\log(X_{0})}{X_{0}} \right)}{X_{0}} \right) \right) \left( \min \left( \frac{2047.69 \log(X_{0})}{X_{0}}, 0.035 \max (28.4336, X_{2}) \right) \right) \] 

\[ X_{DSP41} = (X_{2} \max(X_{1}, \min(2373.23, X_{2})) \max(X_{1}, -\max(X_{2} - 912.142, X_{2})) \] 
\[ \left( \log \left( \frac{-X_{1}}{X_{0}} - X_{0} \right) \right) + X_{1} + 618.377, \frac{X_{2} \max(X_{1}, X_{2}) \min(2373.23, X_{2})}{X_{0}} \] 
\[ - \max(|X_{2}| \left( X_{2} - \log \left( -\sqrt{X_{2}} \right) \right), \min(2373.23, X_{2}) \] 
\[ \log \left( \max \left( \frac{0.0554593}{\sqrt[4]{X_{0}X_{2}}}, \frac{X_{0}}{|X_{2}|} \right) \right) \right) / \left( \log(X_{0}) \right) / X_{0} \] 

\[ X_{DSP42} = \max(X_{0} + 2197.7, X_{1}, (X_{2} \max(X_{2}, \frac{X_{2}^{2}}{X_{0}} - \sqrt{X_{0}} - \max(2X_{0} \vert X_{2} \vert), \] 
\[ - \frac{X_{2} \max(X_{1}, X_{2}) \min(2373.2, X_{2})}{X_{0}} - \min(2373.2, X_{2}) + \frac{X_{2}^{2}}{X_{0}} \] 
\[ + \max(X_{1}, \sqrt[4]{X_{0} - X_{2} - 2718}) \left( -X_{1} - 618.38 \right) \) \] 
\[ / X_{0} + \max(|X_{2}|, \sqrt{-X_{0} - X_{2} - 2718}) \log \left( -X_{1}, \min(\frac{X_{2}}{X_{0}}, X_{0}) \right) \] 
\[ \] 
\[ \] 
\[ \log(\max(X_{1}, X_{2} \frac{X_{2}^{2}}{X_{0}})) \) \] 
\[ \] 
\[ - \max(X_{0} \vert X_{2} \vert, \frac{X_{2} \max(-X_{1}, \log(|X_{1}|))}{X_{0}} - X_{1} \) \]
\[ X_{\text{RSPA1}} = \left( \min \left( \frac{\max \left( X_0, X_1 \right)}{\sqrt{|X_2|}} + X_0 X_2, (\max(X_0, |X_2|) \right) - \max\left( \frac{0.0005 X_1}{\min \left( X_2 - X_0, |X_1| \right)}, \left( X_1 + X_2 + \sqrt{|X_2|} \right) \right) \right) \]

\[ = (0.00014 X_2^{3/2} + X_1 + \min(X_2 - X_0, |X_2|) \max(33.49, \max(0.00016 X_0, |\log(X_0)|) \min(6962.99, \frac{X_0 \max \left( 0.00014 X_2, -\sqrt{|X_2|} \right)}{|X_2| - X_0}) \right) \]

\[ X_0 - X_0 + X_2 + (\min(0.99 X_1, X_1 - X_2) - \log(X_0) \frac{1}{\sqrt{X_2}}) \min(\max(X_2, X_2 - X_0, 3.8 \sqrt{|X_2|}), \max(|X_2|, (X_1 + X_2 + \sqrt{|X_2|})(X_1 + \frac{X_2 - X_0}{X_0} - \log((6224 X_0 + X_1 - \log(-X_0) - 3227)/(\max(14 X_2, 10^5 \frac{X_2}{X_2 - 802.8}))) \max(X_0 + X_1 + \min(X_2 - X_0, 56.3 |X_0|) + \sqrt{|X_2|})(X_1 + \max(X_1, X_1 - X_2) + \min(X_2 - X_0, 6224 X_2 - \log(X_2)))) \min(X_2 - X_0, \min(X_2 - X_0, 3.8 \sqrt{|X_2|})) \right) \]
\[
\frac{6223X_2 - \max (X_0, X_2 - X_0)}{|X_2 - X_0| \max (48.4, \left(X_1 + X_2 + \sqrt{|X_2|}\right) \left(-X_0 + X_1 + X_2 + \frac{\max (X_0, X_1)}{\sqrt{|X_2|}} - 3.8\right))}^{\frac{1}{2}} \]
\[
\sqrt{\min(\log (X_1 - 3223), \frac{6224X_2 - \max (\sqrt{X_2} - 0.00014X_2)}{-X_0X_2 + \sqrt{X_2} - \max (8537, X_2)} + \min (-0.51, X_0X_1)} + \min (-0.51, X_0\min (33.5, -X_0\min (X_1, X_2))) + \min (X_2 - X_0, \sqrt{|X_2|} \log (X_0))] / (\min (\max (8536.91, X_2) + 0.00014X_2, (-\max (X_0, X_2) \log (X_0)) | \max ((X_2 - X_0) \left(X_1 + X_2 + \sqrt{|X_2|}\right), \frac{0.25\max (0.99, X_1 - X_2)}{\sqrt{|X_2|}})])
\[
- X_2^2 + 6224X_2) / ((\min (33.4944, X_2 - X_0)| | \max (0.59, (X_0 + X_1), \min (|X_2|, |\max (3.88\sqrt{|X_2|}, \max (\sqrt{|X_2|} + 33.49)))) (0.00014X_2)^{\frac{1}{2}} + X_1 + | \min (|X_2|, |\min (3.88\sqrt{|X_2|}, \max (\sqrt{|X_2|} + 33.49))) (X_0 / (X_0 + X_1 - X_2) - \max (X_1 - X_2, 0.00014X_2) | - \sqrt{X_2} + 0.59), | \log (X_0) | (-\min (6224.24X_0 + X_1 - \log (X_0))
\[
- X_2^2 \max (0.00014X_2, \sqrt{|X_2|} - X_2)
\]
\[
-3226.93, \frac{|X_2| - X_0}{X_0 - X_0 + X_2 + \sqrt{\max (0.99X_1, X_1 - X_2) | | \max (X_0, \sqrt{X_2} (X_2 - \sqrt{|X_2|}))}) | + \sqrt{X_2} (-X_0 + X_1 + \frac{X_2}{X_2 - 802.9} + X_2 - \log (X_0 + X_1)))) / (\min (33.5, X_2 - X_0, (6224.24X_2 - X_1) / (-\max (8536.91, X_2) + \min (0.6, -X_0\sqrt{X_2} - X_0X_2)))^{\frac{1}{2}} + \min (X_0 + X_1 - \log (-6223.24X_2), X_0 \min (X_1, \log (-6223.24X_2)))^{\frac{1}{2}}))^{\frac{1}{2}}
\]
\[
X_{RSPA2} = \max (\left(\sqrt{|X_2|} + X_1 + X_2\right) - X_0 + X_1 - 0.000143616X_2^2 + X_2), \left|\log (0.00014X_2) - \log (-X_0)\right|)
\]
\[
(6.14)
\]
\[
(6.15)
\]
USA model

\[ X_{CUSA} = \max\left(-\frac{X_0}{X_1^2}, \max(X_1, |\max(X_2, \max(X_0 - X_2, X_2))\right) \]

\[ (2\log\left(\frac{X_0}{X_1^2}\right) + \log((\max(X_1, X_2, \max(7.7884, X_2)) \]

\[ (\log\left(\frac{2271.39X_1^2}{\sqrt{X_0X_1X_2}}\right) + \log\left(\frac{X_0}{X_1(X_0 - X_2 + \log(X_1))}\right) \]

\[ + \min\left(-X_0, \sqrt{\max(X_0, X_1)}\right) + \]

\[ (\max(X_2, \left[|X_0 - X_2 + \log(X_1)| + 2\log\left(\frac{X_0}{X_1^2}\right)\right]) \]

\[ \max(X_2, \log(\max(X_2, X_1))(\log\left(\frac{X_0}{X_1^2}\right) + \]

\[ \log\left(\frac{X_0 - X_2 + \log(X_1)}{X_1^2}\right) + \log\left(\frac{X_0}{X_1^2}\right) \right) \]

\[ (\max(X_2, \log\left(\frac{X_0}{X_1^2}\right), \log(X_1)) \max\left(X_1 + \sqrt{X_2}, X_2\right)\right)^{1/2} + 2.05)))\), \]

\[ (0.00012X_1 + \max(-X_1, (5337.03X_1 + X_2)(0.00012X_1 + \log(X_2)))) \]

\[ (X_2 + \min(-385928., \max(-X_2, \log(X_1) - 7.8))))^{1/2})^{1/2}) / \]

\[ (X_2 - 2412.46) + (\max(X_2, (|X_0 - X_2 + \log(X_1)| \]

\[ + 2\log\left(\frac{X_0}{X_1^2}\right))\max(7.8, X_2), (5337.03X_1 + X_2) \]

\[ \max(X_2, \log(X_0))^{-3}))^{1/2} - X_0 + X_2) \]

\[ (6.16) \]
\( X_{CUS42} = \max(|X_2|, \max(X_2, \max(X_1, (\max(X_2, (X_2 - 2412.46)))) \\
(X_0 - X_2 + \log(X_1))^2, |X_2| (\log\left(\frac{X_0}{X_1^2}\right) + \log\left(|X_0 - X_2 + \log(X_1)| + \log\left(\frac{X_0}{X_1^2}\right)\right) / (X_1^2) + \log(X_0 / (X_2 + 5337.03 \log\left(\frac{\max(X_0, X_1)}{X_1^2}\right)))) + (\max(X_2, (5337.03 X_1 + X_2) \sqrt{X_2 \log(X_0)}, (2 \log\left(\frac{X_0}{X_1^2}\right)) + \log(X_1) + (\max(X_2, \left|X_0 - X_2 + \log(X_1)\right| + 2 \log\left(\frac{X_0}{X_1^2}\right)) \max(7.8, X_2), (0.00012 X_1 + \log(X_2)) \max(X_2, \left|X_0 - X_2 + \log(X_2)\right|) \\
+ \log\left(\frac{|X_0|}{X_1^2}\right) + \log\left(\frac{\min\left(\sqrt{-X_1}, X_0 + 2 \times 10^{-4} \log\left(\frac{65.86}{X_1^{3/2}}\right) + X_2 \log(\log(X_2))\right)}{X_1^2}\right)\right) / (X_1^2)^{1/2} + \log((\min(\sqrt{-X_1}, |X_0 - X_2 + \log(X_1)| + \log\left(\frac{82.7 \sqrt{X_0 \left(\sqrt{X_1} + 64.63\right)}}{X_2}\right)) + \min\left(\sqrt{X_1}, X_0 + 0.00023 \log\left(\frac{\sqrt{X_0 X_1}}{X_1^2}\right) + X_2 \log(\log(X_2))\right) \log(\frac{\min(4662.37, X_0)}{X_1^2}) \max\left(X_1 + \sqrt{X_2, X_2}\right))^{1/2})^{1/2} + X_2) \\
((\max(X_2, \max(7.8, X_2))}
(\log((\min(\sqrt{X_1}, 0.0002 \log\left(\frac{\sqrt{X_0X_1}}{X_1^2}\right) + X_0 + X_2 \log\left(\log \left(\log X_1 \right)\right) / (X_1^2) + \\
\log\left(\frac{|\log \left(\log X_1\right) - X_0|}{X_1^2}\right) + |X_0 - X_2 + \log (X_1)|, \max(7.8, X_2))(\max (X_2, \max(7.8, X_2))(|X_0 - X_2 + \log (X_1)| + 2 \log\left(\frac{X_0}{X_1^2}\right), \max (X_2, \log (X_0)))) \\
(5337.03 X_1 + X_2) \max(\log (\sqrt{X_2} - 6118.39), \max (6884.48, -\frac{X_2}{\min(-5951.71, X_1)}), \min(\log(\max(X_2, \min(3002.14, 622.313 - X_1))), \max(\frac{X_1 - X_0}{X_1 + X_2})) \\
\sqrt{|X_0|, \log(\log(X_1))} \max(3876.54, -X_2)))^{\frac{1}{2}} + \log((\max (X_2, \frac{\sqrt{X_0X_1}}{X_1^2}, \max(X_0, \frac{2271.39}{X_1^2 X_2})(|X_0 - X_2 + \log(X_1)| + \log(X_0), \log\left(\frac{X_0}{X_1^2}\right), \log\left(\frac{2271.39}{X_1^2 X_2}\right)))^{\frac{1}{2}} \\
+ \log\left(\frac{|X_0 - X_2 + \log(X_1)|}{X_1^2} + 2 \log\left(\frac{X_0}{X_1^2}\right)\right))^{\frac{1}{2}} + \log\left(\frac{X_0}{X_1^2}\right)\right)

X_{DUSAI} = (X_0 - X_2)(\log(X_1) - |X_2|) (|X_0 - |X_2||)(\log(2 X_0) - \\
\left|\frac{(X_0 - X_2) X_2 \min(X_0, X_2)}{X_0^2}\right|) / (X_0) | \\
|X_0 - X_2)(\log(X_1) - \log(\log(X_2 - 3853.45)) - \\
\log(X_1) - \frac{|X_2 - 3853.45|}{X_0} \\
(X_0 - X_2) X_2 \min(X_0, \frac{X_0}{2 X_0}) | |)} \\
X_0^2
\[ X_{RU,NI} = \max\{-568.5, (X_2 \log(\min(-1 | X_2 \max(-568.5, (X_2 \max(-568.5, (X_2 \max(-568.5, (X_2 \log(| \min(77.2, -1 | X_2 \max(-568.5, -X_2 \max(-568.5, (X_2 \log(\sqrt{X_2})) / ((X_2 - 77.2)(-\min\left(\sqrt{X_2}, -X_2\right) - 77.2))))) / (-\min(X_2, \sqrt{X_2}) - 77.2))))) / (X_2 - 77.2))\right)\} \]

\[
\max(-568.5, \frac{X_2 \log\left(\sqrt{X_2}\right)}{(X_2 - 77.2)^2})
\]

\[
\min(77.2, -X_2, \max(-1415.18, X_0) \max(-568.4, \\
\max\left(-568.4, \sqrt{X_2}\right) / \min(-X_2, |X_2|) - 77.2\right))(X_2 - 77.2)(-\min\left(\sqrt{X_2}, -X_2\right) - 77.2))))) / (X_2 - 77.2)\)
\]

\[
(-\min(-X_2, |X_2|) - 77.2)) / ((-\min(77.2, -X_2) - 77.2))(-\min\left(\sqrt{X_2}, -X_2\right) - 77.2)) / (X_2 - 77.2)
\]

\[
(-\min(5961.23, -X_2) - 77.2)(\min\left(\sqrt{X_2}, -X_2\right) - 77.2)) / (X_2 - 77.2))\max(-568.5, \left(\min(-X_2, \sqrt{X_2}) + 77.2\right)\right) \]

\[
X_2 \log(| \max(-568.5, \left(77.2 - X_2\right)(\min(-X_2, \sqrt{X_2}) + 77.2)\right)\} \]

\[
77.2 \max(-568.5, \frac{4.35X_2}{X_2 - 77.2}) - \min(77.2, -X_2, \max(-1415.18, X_0) \max(-568.5, \sqrt{X_2} / \min(-X_2, X_2) - 77.2) - 77.2 \min(96.883, X_0) / ((77.2, -X_2)(-\min(-X_2, |X_2|) - 77.2)) \right) \]

\[
\log(| \max(-568.5, \min\left(\sqrt{X_2}, -X_2\right) - \sqrt{X_2}))) / (X_2 - 77.2))\right)
\]

Global model
\[ X_{\text{GLOB1}} = \left( (-X_2) \max (1, |X_2|) \right) \left( (-X_2) \max (0, \max (X_2, 0.00053 |X_2| |X_2|) \right) \]
\[
X_2 \sqrt{\max \left( 0, |X_2| \right)} \frac{1}{|X_2|} \max \left( -X_2, |X_2| \right) / \left( \sqrt{\max \left( 0, |X_2| \right)} \right) \]
\[
\frac{1}{\sqrt{\max \left( 0, |X_2| \right)}} \max \left( -X_2, |X_2| \right) / \left( \sqrt{\max \left( 0, |X_2| \right)} \right) \]
\[
\max \left( -0.00053 X^2_2, |X_2| \right) \frac{1}{|X_2|} \max \left( 1, |X_2| \right) \]
\[
1884.58 \frac{\sqrt{X_2 |X_2| \max \left( 0, |X_2| \right)}}{|X_2|} \left( \sqrt{X_2 (-X_2) \max \left( |X_2|, 0.00056 \Xi \right)} \right) \]
\[
0.00056 \Xi |X_2|^{15/32} \left( \left( \left( \frac{1}{|X_2|} \right) \right) \right) \frac{1}{|X_2|} \max \left( -X_2, |X_2| \right) \]
\[
\max \left( -0.00053 X^2_2, |X_2| \right) \max \left( 1, \max (X_2, 0.00053 X_2 \max \left( |X_2|, 0.00056 \Xi \right)) \right) \]
\[
\frac{2.8265 \sqrt{X^2_2 \sqrt{X_2^{15/2} \max \left( 1, |X_2| \right)}} \max \left( -X_2, |X_2| \right)}{|X_2|^2} \]
\[
\max \left( 1, |X_2| \right) \left( \left( \frac{1}{|X_2|} \right) \right) \left( \frac{1}{|X_2|} \right) \left( \frac{1}{|X_2|} \right) \left( \frac{1}{|X_2|} \right) \]
\[
X_{\text{GLOB2}} = \sqrt{X_2 \max \left( -X_2, X_2 \right) \sqrt{\max \left( 1, |X_2| \right)} \max \left( \left| X_2 \right|, 0.00053 X_2 |X_2| \right)} \]

\[ (6.20) \]
\[ (6.21) \]
\[ X_{RGLOB} = \max(|X_2|, 0.000530622 X_2 \sqrt{|X_2|})(|X_2| \sqrt{|X_2|} \max(1, \max(0, |X_2|))) \]
\[
\max(|X_2|, (2.56(|X_2|^2) \max(|X_2|), 1.68122(|X_2|)(3226.81 - X_2)) \]
\[
\sqrt{|X_2|} \max(1, |X_2|) \left( \frac{1}{2} \max(-X_2, |X_2|) \right) \left( \frac{1}{2} \max(4073.68, \frac{1884.58}{X_2}) \right)^{\frac{1}{2}} \]
\[
\max(0.00053 X_2 \sqrt{|X_2|}, -(3.08232 \sqrt{|X_2|} \max(1, \max(0, |X_2|))) \sqrt{|X_2| \frac{X_2^{3/2}}{\max(-X_2, |X_2|)} \]
\[
\max\left(\frac{1}{|X_2|^2}, |X_2|\right)/(|X_2|^2) \max(4073.68, -X_2))^{\frac{1}{2}} \]
\[
\max(-0.00053 X_2^2, |X_2|) \left( \frac{1}{2} \max(1, \max(X_2, 2.2 X_2)) \right)^{\frac{1}{2}}, \]
\[
\left| \max(1, \max(0, |X_2|)) \right| \left( \frac{1}{2} \max(-X_2, |X_2|) \right) \left( \frac{1}{2} \max(-X_2, |X_2|) \right) \left( \frac{1}{2} \max(-X_2, |X_2|) \right) \]
\]