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

21

Time series analysis of the COVID-19 pandemic in Australia using genetic programming

Rohit Salgotra[1], Amir H. Gandomi[2]

[1]DEPARTMENT OF ECE, THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY, PATIALA, PUNJAB, INDIA; [2]FACULTY OF ENGINEERING & INFORMATION TECHNOLOGY, UNIVERSITY OF TECHNOLOGY SYDNEY, ULTIMO, NSW, AUSTRALIA

1. Introduction

The coronavirus disease 2019 commonly referred to as COVID-19 is an ongoing pandemic, caused by the Severe Acute Respiratory Syndrome Corona virus 2 (SARS-CoV-2). The virus first came into existence on December 8, 2019 in Wuhan, China [1]. On January 9, 2020, the first death was reported and WHO confirmed that a new coronavirus has been identified and isolated [2]. By January 13, 2020, the virus migrated from mainland China to other world and first case outside China was reported in Thailand [3]. The prevention trend started lately and strict actions were imposed in China including public transportation suspension, airport closure, railway line interruptions, closure of highways and state road, shops, public gathering, mass activities, and all other measures were employed that can reduce the community transmission of the disease [4].

Despite such efforts, the virus was not controlled and several cases from different corners of the world were reported by 19 January [3]. On January 31, 2020, the WHO declared it as a global emergency and by March 11, 2020, the COVID-19 was declared as a global pandemic [4]. As of April 25, 2020, 2.5 million people have been affected by the virus with a total death count of more than 150,000 around the globe [5]. The disease that started from a single human being, moved forward to cluster level and now increasing enormously as a community transmission agent across more than 180 countries of the world [4]. The potential effects of COVID-19 have started a number of epidemiological studies on the characteristics of the virus and enormous studies have been conducted to find a possible vaccine for the cure. The International Air Transport Association travel data were used to identify countries where the transmission has spread outside China and also to check the level of infectious disease vulnerability indexes (IDVIs) [6]. The IDVI lies in the range of [0 1], where a higher score means there is lower vulnerability.

Data Science for COVID-19. https://doi.org/10.1016/B978-0-12-824536-1.00036-8

The initial top affected destinations were Bangkok, Hong Kong, Taipei, and Tokyo, all having an IDVI more than 0.65 [6].

As of April 25, 2020, USA is the most affected country with a total of 830,053 cases and Spain being the second most affected area with 213,024 cases. The other countries where the count is enormously increasing are Italy (189,973), Germany (150,383), and UK (138,082). As a global pandemic, the scale of COVID-19 has grown from some few numbers to several folds of magnitude in a matter of weeks and in some cases from hundreds to thousands in couple of days. As already studied, the growth rate of pandemic ranges from 0.2 to 0.3, that is, a daily increase of 20% to 30% in new cases [7]. This is evident from China, France, Germany, UK, Spain, and Italy. But in the case of USA, the increase rate is much more [4]. Average estimates like this can help researchers to design and calibrate disease transmission models, before further investigation and intervention policies of the possible effects of the pandemic [8].

As far as Australia is concerned, total cases so far is 6667 with a confirmed death count of 76 [4]. The transmission classification category is still cluster level or it can be said that it is the second transmission stage. Most of the third world countries have already crossed third and fourth stages (community level transmission) but due to the continuous efforts of the Australia authority, the virus is still under check. The major concern about this infectious transmitting virus is that it has shown adverse effects on people of elder age and those who are already suffering from some sort of heart or respiratory ailments [9]. As the Australian population is grown up and majority of the people are of old age, it is a big concern for the authorities to keep a check on the virus, so that it may not transmit from cluster to community level.

Thus it becomes essential to further estimate the total number of infections in near future to analyze the spread of the disease. To that end, various mobility models have been used by the research community to obtain comparable numbers, and various reports have been published for different countries of the world. For China, where the pandemic started, it was estimated that with in a span of two to three days, the virus has the capability to increase 10% to 15% [10]. The major studies include Weibull distribution−based model [11], stochastic simulations [12], lognormal distribution [13], exponential growth, maximum likelihood estimation [14], and others [15]. Though none of the methods could estimate the exact reproduction rate but the average incubation rate was reported as 5.1 days [11].

In present work, genetic programming (GP) [16] modeling has been used to estimate the possible spread of COVID-19 in Australia. GP is an enhanced version of genetic algorithm [17], in which solutions are computer programs instead of binary strings [18]. More precisely a recent extension of GP commonly referred as gene expression programming (GEP) [19] has been analyzed to build a predictive model for the total number of confirmed cases (CCs) and death cases (DCs) of COVID-19 in Australia. GEP approaches are more efficient and can be used as an alternative to classical techniques. A major advantage of using GEP over the conventional methods and artificial neural network is its stability to generate simple prediction equations. Also, the GEP does not

need any prior relationship to develop prediction model. Numerous researchers have employed the GEP models to discover complex environments and derive prediction-based models [20,21]. The newly proposed model is based on the raw data on the total number of CC and DC from the WHO situation reports (updated daily since January 31, 2020).

2. Technical preliminaries and model calibration

In present work, a highly effective evolutionary algorithm namely GEP has been used for high resolution CC and DC-based pandemic modeling in Australia. Various other methods such as Australian Census-based Epidemic Model (AceMod) and others have been previously used and validated for simulation of pandemic influenza in context with Australia [22]. The same AceMod has also been used to simulate the COVID-19 patterns in Ref. [22]. This method uses a discrete stochastic agent-based model to understand and investigate the complex outbreak of COVID-19 scenarios across the country. But such kind of modeling is classical and requires much more data and simulation scenarios to be performed to predict the actual outcome. Also they can be used to simulate the current environment but pose very challenging implementation when compared with their counterparts. The GEP-based model on the other hand is very simple and can be calibrated easily. These models even predict viable solutions under minimal constraints and maximum accuracy [20]. The experimental tests performed using the CC and DC cases for Australia from the date of first outbreak to current scenario. A detailed methodology used for GEP modeling for the COVID-19 is presented in the subsequent subsection.

2.1 Gene expression programming

GP is an enhancement of GA and is based on the Darwinian theory of natural selection. GP creates computer-based programming equations or data to find a relationship between the input and the output parameters [20]. GP in general is a computer-based program that is simulated in the form of a tree structure and declared in a functional programming language [16]. In this kind of setup, GP consists of a hierarchical structure with terminals and functions [20]. The current version of GP is the GEP which was first developed by Ferreira et al. [19] and consists of five major components. These include function set, terminal set, control or tuning parameters, fitness function, and terminal condition. The GEP uses fixed character length strings instead of conventional tree representation of GP and are subsequently expressed as para trees commonly called as expression trees (ETs). Here, main advantage of this kind of strategy is that it is extremely simple and works at chromosome level. Also because of its multigenic properties, it can be used for evolution of more nonlinear and complex programming composed of several sub-programs [23]. Each GEP consists of symbols having fixed length and comprises of terminal set (e.g., a, b, c, 6) and function set (e.g., -, +, /, Log, $\times$). Thus in terms of both terminal set and function set, a GEP can be having multiple chromosomes which are capable of representation in the form of any parse tree. To decode this information in

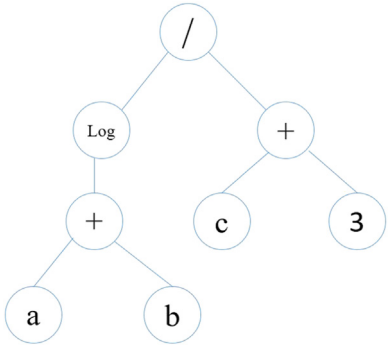


FIGURE 21.1 Representation of an expression tree.

this chromosome, Karva language is used [24]. A typical gene generated using GEP in Karva language is given by

$$/Log++c3ab \tag{21.1}$$

where a, b, and c are variables and 3 is a constant value. The expression in Eq. (21.1) is called as a K-expression or generally a Karva notation. The model thus formulated can be evolved in the form of an ETs. A simplified structure based on the above discussed problem is given by Fig. 21.1.

The expression in Eq. (21.1) is converted into k-expression from the first position which is basically the root of ET, and reads through the model from functional nodes to the terminal node. This type of representation allows for a more complex and quicker understanding of the mathematical intricacies [25]. The k-expression thus formulated in the form of mathematical equation and is given by

$$Log(a+b)/(c+3) \tag{21.2}$$

Overall it should be noted from the above k-expression that the length of genes in a GEP remains same whereas the number of ETs vary with respect to the problem complexity. This further signifies that there are certain elements which can not be used for genetic mapping. So for a GEP to be efficient, the generic length for any k-expression should be less than or equal to the total length of a GEP gene. Here it should be noted that a GEP employs a trial-error method to randomly select a genome. The head consists of both terminal and function symbols whereas tail has only the terminal symbol [19].

The GEP algorithm initiates with a random initialization of fixed length-based chromosome for each member from the whole set of population. The second step is to evaluate the chromosomes, evaluate the solutions, and finally select the best fit solution based on the fitness of respective individuals to reproduce with modifications. All of this is followed for some predefined set of generations or unless and until the termination criteria is met. The schematic diagram for the fundamental steps of GEP is given by Fig. 21.2. Furthermore, it should be noted that the fitness of these solutions is updated based on Roulette wheel sampling with elitism. Thus helping the algorithm in

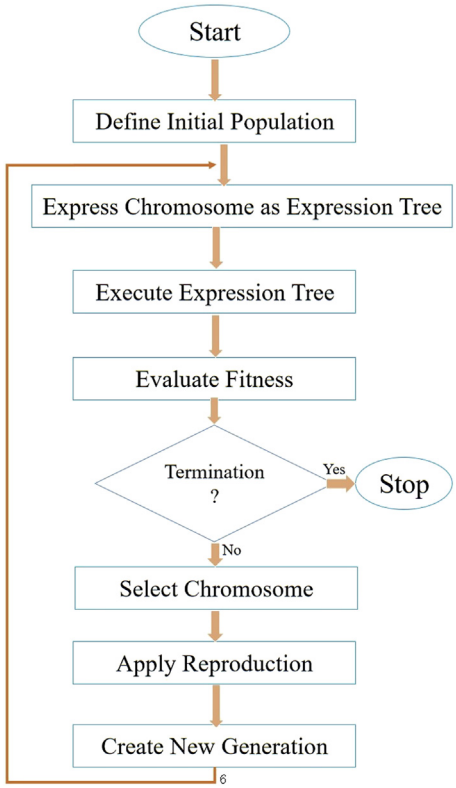


FIGURE 21.2 Representation of a gene expression programming algorithm.

optimizing and cloning the best individual in consecutive generations and ultimately finding the global solution [20].

2.2 Proposed gene expression programming model

To accurately assess the total COVID-19 cases across Australia, the effect of both CC and DC was taken into consideration for model development. Eight former records are considered in the time series models and GEP model selected the best ones out of them. The experimental database was divided into two subsets including training and validation/testing phase. As already known that a single run can not define the proper performance of any meta-heuristic algorithm because of its random nature. In present work, multiple runs of the same experimental data set were performed to decrease the possible error. This property really help when the total number of instances available for the experimental data are not in abundance. Of the total experimental data, 70% of the data was used for training purpose, and the rest were used for validation/testing phase. Note that the training data were used for gene evolution, and the best model was selected based on the correlation coefficient on the training data. Thus a final model whose output performance was better for training but not necessarily for testing.

In addition to this, the parameters of a GEP algorithm also affect the model generalization capability. The parameters of GEP were changed for multiple different runs to find the global optimal solution. The initial selection was made on the basis of previously selected models as suggested by Ref. [19]. To calculate the overall fitness of evolved program, the fitness function defined by Eq. (21.3) is used.

$$Fitness = \left(\frac{1}{1 + MSE}\right) \times 1000 \tag{21.3}$$

where *MSE* is the mean squared error of the evolved program. A detailed parametric study for the presented GEP model is given in Table 21.1.

The GEP algorithm was implemented using GeneXpro tool [25]. For genetic operators, the parameter settings as given in Table 21.1 were used. The algorithm was run for every set of parameter until no desirable improvement can be extracted from the GEP model. The model's architecture as evolved by GEP has been calculated by using head size and total number of genes. Here, number of genes for a single chromosome determines the total number of terms in the model and each gene corresponds to each sub-ET. Four optimal levels were devised for head size and five for the number of genes. If gene size becomes greater than one, the average linking function has been used to link the mathematical model. In this study, simple mathematical functions were taken into consideration to get the optimal GEP models. Furthermore, note that the program was run unless no further improvement in the performance was noticed. A set of statistical parameters of the GEP model is presented in Table 21.2.

Table 21.1 Parameter Settings for gene expression programming algorithm.

Parameter	Settings
General	
Chromosome	30
Gene	5
DC size	5
Head size	4
Tail size	5
Gene size	14
Linking function	Average
Genetic operator	$+, -, \times, \div, \sqrt{}$
Mutation rate	0.00206
Inversion rate	0.00546
IS and RIS transposition rate	0.00546
One-point and two-point recombination rate	0.00277
Gene recombination and transposition rate	0.00277
Numerical constants	
Constant per gene	10
Data type	Floating-point
Range	[−10, 10]

Table 21.2 Statistical Parameters of gene expression programming model for external validation.

Item	Formula	Condition	GEP CC	GEP DC
1	R	$0.8 < R$	0.9998	0.9992
2	$k = \left[\sum_{i=1}^{n}(h_i \times t_i)\right] \Big/ h_i^2$	$0.85 < k < 1.15$	0.9996	0.9994
3	$k' = \left[\sum_{i=1}^{n}(h_i \times t_i)\right] / t_i^2$	$0.85 < k' < 1.15$	1.0000	0.9998
4	$m = \left(R^2 - Ro^2\right)/R^2$	$\lvert m\rvert < 0.1$	−0.00036	−0.00154
5	$n = \left(R^2 - Ro'^2\right)/R^2$	$\lvert n\rvert < 0.1$	−0.00026	−0.00155
6	$R_m = R^2 \times \left(1 - \sqrt{\lvert R^2 - Ro^2\rvert}\right)$	$0.5 < R_m$	0.9837	0.9592
where	$Ro^2 = 1 - \left[\sum_{i=1}^{n}\left(t_i - h_i^0\right)^2\right] \Big/ \left[\sum_{i=1}^{n}\left(t_i - \overline{t_i}\right)^2\right]$	$h_i^0 = k \times t_i$	1.0000	0.9999
	$Ro'^2 = 1 - \left[\sum_{i=1}^{n}\left(h_i - t_i^0\right)^2\right] \Big/ \left[\sum_{i=1}^{n}\left(h_i - \overline{h_i}\right)^2\right]$	$t_i^0 = k' \times h_i$	1.0000	1.0000

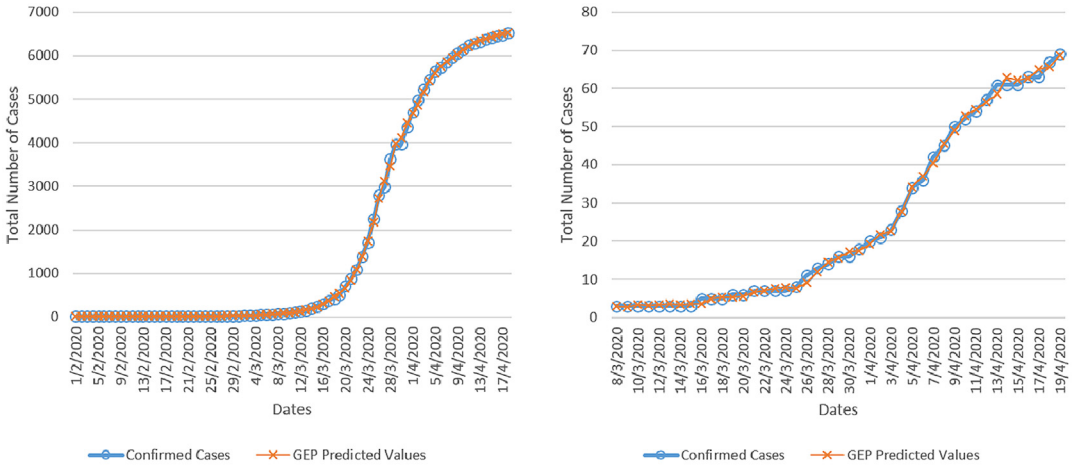
3. Proposed gene expression programming–based formulation for best OBJ

In present work, two different formulations based on the CC and DC are proposed. A comparison of the experimental to predicted values for both the CC and DC cases is given in Fig. 21.3. The above-mentioned mathematical formulas present a complex organization of variable, operators, and constants and are used to predict the output. The ETs for both CC and DC are given by Fig. 21.4, and numerical equations can be derived from them. As given by the figures, it can be seen that the proposed equations are divided into five independent components (genes or simply subprograms) and are consecutively linked by the average function. Each of these subprograms indicate individual aspects of the problem so that meaningful overall solution can be developed [20]. Thus it can be said that each newly evolved sub-function consists of important information about the psychology of the final resultant model. Each gene thus formulated is expressed in the final equation and is responsible for finding a particular facet of the problem. This kind of information is necessary for further evaluations at chromosomal level [25].

4. Model validity and comparative study

The basic metrics for model evaluations are the correlation coefficient (R) and root mean square error ($RMSE$), which are calculated as

$$MAE = \frac{\sqrt{\sum_{i=1}^{n}\left(h_1^2 - t_i^2\right)}}{n} \tag{21.4}$$



(a) Confirmed Cases (b) Death Cases

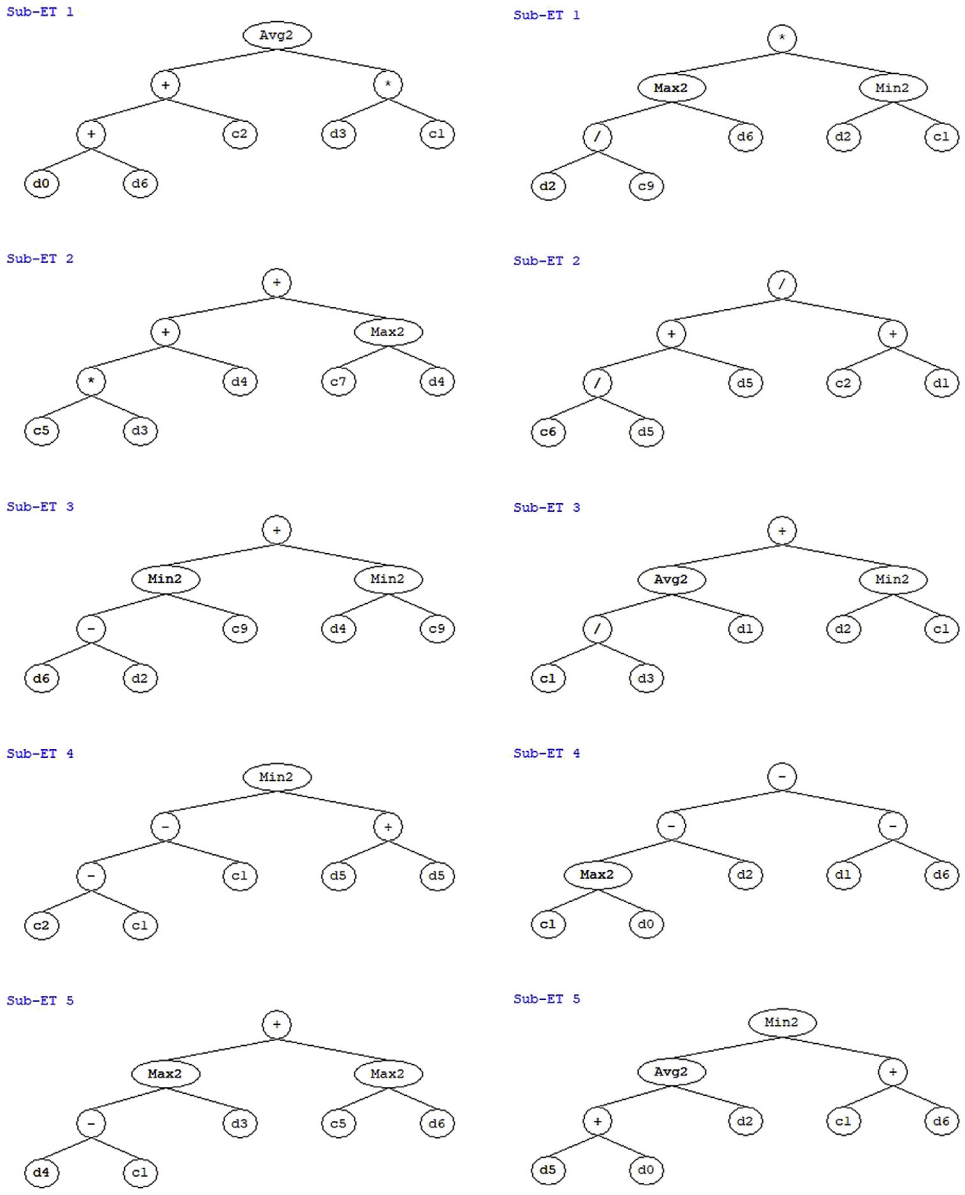
FIGURE 21.3 Experimental versus predicted cases for COVID-19 in Australia using gene expression programming model.

$$RMSE = \frac{\sum_{i=1}^{n} |h_i - t_i|}{n} \tag{21.5}$$

$$R = \frac{\sum_{i=1}^{n} \left(h_i - \overline{h_i}\right)\left(t_i - \overline{t_i}\right)}{\sqrt{\sum_{i=1}^{n} \left(h_i - \overline{h_i}\right)^2 \sum_{i=1}^{n} \left(t_i - \overline{t_i}\right)^2}} \tag{21.6}$$

where n is the maximum number of samples, t_i and h_i are the calculated and actual outputs, $\overline{h_i}$ and $\overline{t_i}$ are averages of the actual and calculated outputs for the ith output. Note that R values alone cannot be considered as a good indicator for evaluating accuracy of any model. The major reason of this is that R values do not change by shifting the output of a predictive model. The other parameters may include an error function such as *RMSE* where a lower value of this function means a more precise model. For any model to be accurate and reliable, Smith et al. [26] stated that a strong correlation must exist between the measured and the predicted values. If a model has a $R > 0.8$, that model is considered as a good model [19]. Overall, any model with low RMSE and high R value has the capability to predict values to a higher acceptable level of accuracy [27]. In present work, the predicted statistics for both CCs and deaths across Australia are given in Table 21.3.

A new criteria for external validation of GEP model was proposed in Ref. [28]. It presented that the regression (k or k') slope should around the origin and must be close to 1. The value of m and n should be lower than 0.1. Another important study states that the value of external predictability of a model that is $R_m > 0.5$ [29]. They further formulated that the squared correlation coefficient (through the origin) (Ro^2) between



(a) ETs for Confirmed Cases (b) ETs for Death Cases

FIGURE 21.4 Expression trees for the modeling of COVID-19 in Australia.

Table 21.3 Overall Performance of gene expression programming model for confirmed case and death case across Australia.

Model ID	RMSE	R
CC	24.5393	0.9997
DC	525.2309	0.9984

the predicted and the experimental values, or the coefficient (Ro'^2) between the experimental and the predicted values must be close to 1 [20]. The conditions for external validation are presented in Table 21.2. The major factors concerning validation phase ensure that the proposed model has a good prediction power and is strongly valid. Taking all of the above points under consideration, it can be said that the proposed model satisfies all the required conditions and hence can be treated as a valid predictive model. Furthermore, it has distinction with respect to conventional models as it can be readily implemented and uses minimal set of initial conditions for implementation.

The GEP approach used in present work is based on the time series data to determine the CC and DC of the model. The models thus can be consecutively used for preliminary design stages [30]. Another important feature of this model is that it can be used to check the general behavior of coronacases across Australia and access the future requirements.

5. Variable importance

The contribution of each and every variable in the GEP model was evaluated using a variable importance of the all the variables in the model [31]. The variable values for both CC and DC cases across Australia are presented in Fig. 21.5. The importance of each variable is found by randomization of input values and then finding the decrease in R^2 between the model predicted output and the target value. The results thus obtained are normalized in such a way that their addition amounts to 1. According to Fig. 21.5, in both models, the values at a week ago ($d6$) are the most important variable and has the most influence on the models. It can be seen that CC is highly sensitive to $d4$ too.

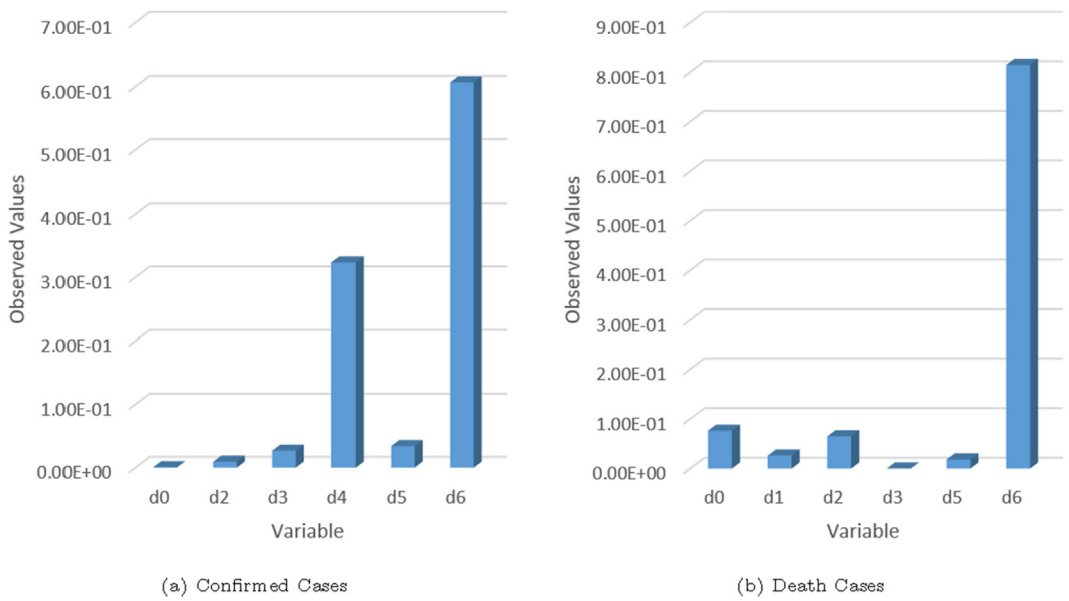


FIGURE 21.5 Contribution of predictor variables for COVID-19 in Australia.

As presented, from the above analysis, it can be said that a GEP-based modeling of COVID-19 provides a very reliable solution. This is because of high-correlation coefficient and lower RMSE. The major reason for this exceptional performance is the simplicity of COVID-19 data sets and high accuracy of the GEP models. Furthermore it should be noted that for simple data sets, GEP models are highly accurate as compared to their ANN and optimization-based counter algorithms. Another important feature of GEP models is that a simple transfer function, relating to the inputs and the outputs, can be derived and further optimized using global algorithm algorithms such as krill herd algorithms [32], naked mole-rat algorithm [33], and others.

6. Conclusion

A robust variant of GEP was used to formulate CC and DCs of COVID-19 in Australia. Two empirical models were derived for the prediction of CC and DC in Australia. The proposed models were developed based on the WHO reports on the total number of CCs and DCs updated on a daily basis since January 31, 2020. The following conclusions have been drawn based on the formulated models:

- The proposed models provide reliable predictions for both CCs and death count. Also, the GEP prediction models proposed, satisfy all the required conditions for external validations.
- The verification of the models were done in term of RMSE and R^2, where a higher value of R^2 close to 1 has been achieved for both CC and DC. Hence further validating the solution quality and higher prediction ability of the proposed models.
- The ETs have been drawn and simpler sophisticated equations can be derived without the requirement of time-consuming laboratory-based implementations for the model. The equations thus derived can be used to optimize the model using different heuristic algorithms such as differential evolution, ant colony, and others.
- Another important observation from the results of variable importance is that both the proposed models are very sensitive to the value in a week before ($d6$), and they are less sensitive to $d0$, $d1$, $d2$, $d3$, and $d5$ in comparison to others.
- The distinctive feature of GEP model which makes it more reliable is that it is based on experimental data and not just assumptions, which are used in conventional models. Also, it can work on lower data and provide reliable predictions. Similarly as more data are added, these models can be significantly improved.

Thus overall we can say that GEP models proposed in present work are highly reliable and can be considered as benchmark for time series predictions. But as the data point increase many fold, they are found to have some limitations. So as a future direction, when more data regarding the COVID-19 become available, new GEP model−based equations can be derived and high-cost evolutionary algorithms can be used for optimization of prediction models.

References

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (2020) 497–506.

[2] WHO, Statement Regarding Cluster of Pneumonia Cases in Wuhan, China, World Health Organization, Geneva, Switzerland, 2020. Available online, https://www.who.int/china/news/detail/09-01-2020-who-statementregarding-cluster-of-pneumonia-cases-in-wuhan-china (Accessed on 17 February 2020).

[3] WHO, Novel Coronavirus–Thailand (ex-China), World Health Organization, Geneva, Switzerland, 2020. Available online, https://www.who.int/csr/don/14-january-2020-novel-coronavirus-thailand-ex-china/en (Accessed on 17 February 2020).

[4] WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 − 11 March 2020, 2020. Online; Accessed 21 March 2020.

[5] WHO. Situation Report-95, World Health Organization, Geneva, Switzerland, 2020. Available online, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200424-sitrep-95-covid-19.pdf.

[6] M. Moore, B. Gelfeld, A.T. Okunogbe, P. Christopher, Identifying Future Disease Hot Spots: Infectious Disease Vulnerability Index, RAND Corporation, Santa Monica, CA, USA, 2016. Available online, https://www.rand.org/pubs/research-reports/RR1605.html (Accessed on 17 February 2020).

[7] J. Riou, C.L. Althaus, Pattern of Early Human-To-Human Transmission of Wuhan 2019-nCoV, bioRciv, 2020.

[8] J.A. Backer, D. Klinkenberg, J. Wallinga, The Incubation Period of 2019-nCoV Infections Among Travellers from Wuhan, China, medRciv, 2020.

[9] S.L. Chang, et al., Modelling Transmission and Control of the COVID-19 Pandemic in Australia, 2020, p. 10218, arXiv preprint arXiv:2003.

[10] J.M. Read, J.R.E. Bridgen, D.A.T. Cummings, A. Ho, C.P. Jewell, Novel Coronavirus 2019-nCoV: Early Estimation of Epidemiological Parameters and Epidemic Predictions, medRciv, 2020.

[11] P. Boldog, et al., Risk assessment of novel coronavirus COVID-19 outbreaks outside China, J. Clin. Med. 9 (2) (2020) 571.

[12] N.M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A.R. Akhmetzhanov, S.-M. Jung, B. Yuan, R. Kinoshita, H. Nishiura, Epidemiological Characteristics of Novel Coronavirus Infection: A Statistical Analysis of Publicly Available Case Data, medRciv, 2020.

[13] Q. Zheng, H. Meredith, K. Grantz, Q. Bi, F. Jones, S. Lauer, JHU IDD Team, Real-time Estimation of the Novel Coronavirus Incubation Time, 2020. Available online, https://github.com/HopkinsIDD/ncov-incubation (Accessed on 17 February 2020).

[14] S. Eubank, H. Guclu, V.A. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang, Modelling disease outbreaks in realistic urban social networks, Nature 429 (6988) (2004) 180–184.

[15] T. Liu, J. Hu, M. Kang, L. Lin, H. Zhong, J. Xiao, A. Deng, Transmission Dynamics of 2019 Novel Coronavirus (2019-nCoV), 2020.

[16] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1992.

[17] D.E. Goldberg, J.H. Holland, Genetic Algorithms and Machine Learning, 1988.

[18] W. Banzhaf, P. Nordin, R. Keller, F. Francone, "Genetic Programming–An introduction." on the Automatic Evolution of Computer Programs and its Application, dpunkt/Morgan Kaufmann, Heidelberg, Germany/San Francisco, 1998.

[19] C. Ferreira, Gene expression programming: a new adaptive algorithm for solving problems, Complex Syst. 13 (2) (2001) 87–129.

[20] A.H. Gandomi, A.H. Alavi, M.R. Mirzahosseini, F.M. Nejad, Nonlinear genetic-based models for prediction of flow number of asphalt mixtures, J. Mater. Civ. Eng. 23 (3) (2011) 248−263.

[21] A.A. Javadi, M. Rezania, Applications of artificial intelligence and data mining techniques in soil modeling, Geomech. Eng. 1 (1) (2009) 53−74.

[22] K.M. Fair, C. Zachreson, M. Prokopenko, Creating a surrogate commuter network from Australian Bureau of Statistics census data, Scient. Data 6 (1) (2019) 1−14.

[23] A.H. Gandomi, S.K. Babanajad, A.H. Alavi, Y. Farnam, Novel approach to strength modeling of concrete under triaxial compression, J. Mater. Civ. Eng. 24 (9) (2012) 1132−1143.

[24] A.H. Alavi, A.H. Gandomi, A robust data mining approach for formulation of geotechnical engineering systems, Eng. Comput. 28 (3−4) (2011) 242−274.

[25] GeneXpro Tools 4.0 [Computer Software], GEPSOFT Ltd, Bristol, UK, 2006.

[26] G.N. Smith, Probability and Statistics in Civil Engineering, Collins, London, 1986.

[27] I.E. Frank, R. Todeschini, The Data Analysis Handbook, Elsevier, Amsterdam, 1994.

[28] A. Golbraikh, A. Tropsha, Beware of q2!, J. Mol. Graph. Model. 20 (4) (2002) 269−276.

[29] P.P. Roy, K. Roy, On some aspects of variable selection for partial least squares regression models, QSAR Comb. Sci. 27 (3) (2008) 302−313.

[30] A.H. Gandomi, A.H. Alavi, G.J. Yun, Nonlinear modeling of shear strength of SFRC beams using linear genetic programming, Struct. Eng. Mech. 38 (1) (2011) 1−25.

[31] A.H. Gandomi, A.H. Alavi, C. Ryan (Eds.), Handbook of Genetic Programming Applications, Springer, Cham, 2015.

[32] A.H. Gandomi, A.H. Alavi, Krill herd: a new bio-inspired optimization algorithm, Commun. Nonlinear Sci. Numer. Simulat. 17 (12) (2012) 4831−4845.

[33] R. Salgotra, U. Singh, The naked mole-rat algorithm, Neural Comput. Appl. 31 (12) (2019) 8837−8857.