Asia-Pacific Journal of Science and Technology

https://www.tci-thaijo.org/index.php/APST/index

Published by the Research and Graduate Studies,
Khon Kaen University, Thailand

The logistic growth regression model with the genetic algorithm for predicting the third wave of the COVID-19 epidemic in Thailand

Rati Wongsathan[1,*]

[1]The Research Institute and Faculty of Engineering and Technology, North-Chiang Mai University, Chiang Mai, Thailand
[*]Corresponding author: rati@northcm.ac.th

Received 12 June 2021
Revised 18 October 2021
Accepted 4 December 2021

Abstract

The world is currently facing the novel coronavirus 2019 (COVID-19). Thailand, with a high basic reproduction number (2.27), the situation remains serious as the disease spreads throughout the country. Applying various control measures to contain the outbreak has increased the need for policymakers to assess the scale of the epidemic. In this study, a logistic growth regression (LGR) model is implemented to characterize the trends and estimate the final size of the third wave of the epidemic in Thailand at both the provincial and national levels. The parameters of the LGR are fine-tuned through the genetic algorithm assisted by the Gauss-Newton algorithm (GA/GNA). The outbreak data from the previous two waves of infection is used to validate the model performance. As a result, the LGR-GA/GNA model provides goodness-of-fit with a low RMSE, high R^2, and highly significant parameters. Furthermore, when compared to the LGR model parameterized by particle swarm optimization and ant colony optimization, the proposed model outperforms the rest. In addition, to verify the prediction performance by comparing with the Susceptible-Infectious-Recovered (SIR) model, the proposed model improves the prediction accuracy better than the other. As the work was completed on May 6, 2021, the study found a possible increasing trend of COVID-19 for some vulnerable provinces and the whole country and an estimated final and peak size of the epidemic and their occurrences. The study concluded that the epidemic size of the third wave of COVID-19 in Thailand was about 190,000 by mid-July 2021.

Keywords: COVID-19, Logistic growth regression, Genetic algorithm, Gauss-Newton algorithm, Particle swarm optimization, Ant colony optimization, SIR model

1. Introduction

The COVID-19 (SARS-CoV-2) pandemic is very severe and life-threatening to the human race in the 21st century. For Thailand, the situation has worsened since the first case of infection was detected on January 12, 2020. The virus's spread seemed very slow for a while until the first super-spreader occurred, after which the infection rate peaked exponentially. The outbreak subsided by May 2020 after lockdown measures had been taken. In December 2020, the second wave of the outbreak occurred, which clustered primarily around large migrant worker communities, but the virus also spread to other provinces. After infections ended in March 2021, due to the restriction of those vulnerable areas, the situation worsened again with a new third wave of infections (April 1, 2021), which spread throughout the country. This spread was partially connected to bar patrons in Bangkok (the capital) and workers returning to their birthplaces during the traditional Thai New Year festival, causing COVID-19 infection cases to drastically increase to more than ten times as much as the same time in the previous year. Thailand currently tops the world rankings with a reproductive number of 2.27, meaning that a COVID-infected person transmits the infection to another 2.27 persons, whereas other places keep this rate low, including India (1.49), USA (1.08), and Brazil (1.10) (April 13, 2021) [1]. In at least 62 of Thailand's 77 provinces where COVID-19 is spreading, a number of patients with mild symptoms are being treated at the special field hospital to avoid hospitalization. However, in the absence of timely and accurate estimates of infected cases, provincial governors cannot provide sufficient facilities in advance. Health care providers may make mistakes in resource planning. Therefore, in short-term and long-term planning to control infectious transmission, a predictive model must be used to determine the trajectory and the final size of the COVID-19 epidemic, as well as the practical measures. Furthermore, it also helps the disease control authority systematically respond to a future surge in infections. However, this study is limited specifically to the context of Thailand.

In the literature survey, the characteristics of the first two waves of the COVID-19 epidemic in Thailand were illustrated using a linear regression model that incorporated the ambient temperature [2]. However, the ambient temperature is very different throughout the country. In addition, deterministic-mechanistic models were applied in predicting the second wave of the epidemic [2-3]. In contrast, studies of predictive models in other countries are extensive and advanced in recent years. These include a mechanistic-based Susceptible-Infectious-Recovered (SIR) model [4-6] and its derivatives, such as, the S-Exposed-IR (SEIR) model [7] and the SIR/Death (SIRD) model [8], and a statistics-based model, such as the multiple linear regression (MLR) [9], the nonlinear regression model using Bayesian approach [10] and the autoregressive integrated moving average (ARIMA) [11]. For a modern data-driven technique, the machine learning-based artificial intelligence approach has been increasingly used in the field of prediction, such as predicting the number of daily cases of COVID-19 using a neural network (NN) model [12]. Additionally, the model-free methods, such as logistic growth function, hyperbolic tangent function, and Gaussian function, were proposed by employing curve fitting [13-14].

However, the SIR model and its variant are based on the strongly idealistic assumptions that the number of people always remains constant or the individuals that recover have immunity, which may not be valid and have not yet been proven in the case of COVID-19 epidemic. Further, the MLR and ARIMA approaches require multi-step prediction, i.e., prior investigation of the stationary of the data, data transformation, lagged identification, parameter estimation, selection of the best model, and diagnostic to check the correlation of residuals. Furthermore, the NN model requires a number of datasets to train the network, otherwise, an over-fitting is likely to occur. To overcome the problem, a long-short-term-memory (LSTM) in the field of deep learning integrated the complete data of the SARS epidemic (in 2003) into the COVID-19 data in the training step [7]. However, SARS and COVID-19 are different, and the SARS data is not up to date. Therefore, machine learning to produce a COVID-19 predictive model, may be inaccurate and un-reliable [14].

Typically, the key to accurately predict an epidemic is parameter estimation of the model. There are various methods of parameter estimation used in COVID-19 epidemic models. The ordinary least squares method was used to evaluate the unknown parameters of the MLR models [9,15]. Gradient-based search methods include the Gauss-Newton algorithm (GNA) used to iteratively solve the parameters of the logistic growth regression (LGR) model [13], the Gompertz, the von Bertalanffy, and the cubic polynomial models [16]. Moreover, the Levenberg-Marquardt algorithm (LMA) combining the GNA, and gradient descent method was used to fit the nonlinear SIR model [4]. On the other hand, the non-gradient search method, for example, the Nelder-Mead (NM) simplex method was applied to estimate the parameters of the Gompertz model [17]. In addition, for global optimization using stochastic search, simulated annealing (SA) was employed to optimize the parameters of the SEIR model [18]. Furthermore, particle swarm optimization (PSO) - a bio-inspired computation is used to identify the parameters of the SIR and SEIR model [19-20].

In the early stages of the COVID-19 epidemic with existing in a short period, this work presents a simple but powerful predictive model based on LGR method. Moreover, genetic algorithm (GA) as a global search technique assisted by the GNA, a fast local convergence, is applied to fine-tune the LGR's parameters for enhancing prediction performance. The proposed LGR-GA/GNR model is used to extrapolate the trends and estimate the epidemic size and cessation date of the ongoing third wave of the COVID-19 epidemic in Thailand at both provincial and national levels. To validate that the proposed model can be used as a predictor of the ongoing outbreak it is first conducted to project the trend in the first and second waves of the COVID-19 epidemic that have passed their peak and completed their cycle. Moreover, this study also conducts a performance comparison of the parameter estimation between the GA and the two well-known population-based search methods, PSO and ant colony optimization (ACO). Furthermore, the performances of prediction of the proposed model are compared with those of the deterministic SIR model. In the test, 102 days of data (February 9, 2020 - May 20, 2020) and 101 days of data (December 11, 2020 - March 31, 2021) of daily infected cases for the first and second waves of the epidemic, respectively, are divided into two parts of training/testing of different sliding windows each to formulate/validate the LGR model. On the other hand, the predicted target focuses on the ongoing third wave of the epidemic, for which 36 days (April 1, 2021 - May 6, 2021) of data for daily infected cases have been collected.

2. Materials and methods

2.1 LGR model and GNA and GA Optimization

In mathematical epidemiology, the logistic growth model is basically used to model the spread of disease, including COVID-19. A general logistic differential equation is expressed by.

$$\frac{\mathrm{dC(t)}}{\mathrm{dt}} = \mathrm{RC(t)}\left(1 - \frac{\mathrm{C(t)}}{\mathrm{K}}\right) \tag{1}$$

where C(t) is the number of accumulative COVID-19 cases as a function of time t (in units of days), R is the logistic growth rate or infection rate, and K is called carrying capacity, or final epidemic size.

In fact, Equation (1) is a Bernoulli equaion which yields the logistic function solution as follows:

$$\mathrm{C(t)} = \frac{\mathrm{K}}{1+\mathrm{A}\exp(-\mathrm{Rt})} \tag{2}$$

where $\mathrm{A} = \frac{\mathrm{K}-\mathrm{C}(0)}{\mathrm{C}(0)}$ is a constant, and C(0) is the initial number of cases.

In general, the logistic function (S-curve) is exponential increasing rapidly for some initial number of times, but the growth decreases slowly to a certain limit below the horizontal asymptote when it passed the turning point (ln (A)/R, K/2) or the peak of the pandemic. This function is used to directly model the patient's cumulative cases. Equation (2) can be expressed in LGR form as follows:

$$\mathrm{C(t)} = \hat{\mathrm{C}}(\mathrm{t};\theta) + \varepsilon(\mathrm{t}) \tag{3}$$

where $\hat{\mathrm{C}}(\mathrm{t};\theta)$ is the logistic growth model, where $\theta = \{\mathrm{K, R, A}\}$ is the set of parameters, and ε(t) is the residuals assumed to be independent and identically distributed random variables.

In this work, the θ is estimated by two different methods, one is a gradient-based NLS method of GNA-a classical iterative technique by minimizing the sum of the residual squares, Equation (4), and the other is the evolutionary search-based GA as a gradient-free method by maximizing the fitness function, Equation (5).

$$J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \sum_t \left(C(t) - \hat{C}(t:\boldsymbol{\theta})\right)^2 \tag{4}$$

$$F = \max_{\boldsymbol{\theta}} \frac{1}{1+0.1\times\sum_t^{\lceil 0.1m\rceil} | C(t) - \hat{C}(t;\boldsymbol{\theta}) | +0.9\times\sum_{\lceil 0.1m\rceil+1}^{m} | C(t) - \hat{C}(t;\boldsymbol{\theta}) |} \tag{5}$$

where t =1, 2, …, m ≥ 3, m is a sampling size, and $\lceil\cdot\rceil$ denotes the rounding up to a positive integer. The second and third terms of the denominator in Equation (5) are weighted residuals that occur at the flattening and nonlinear high slope portions of the logistic function.

For the GNA, the minimum value of J occurs when the gradient $\partial\mathrm{J}/\partial\theta_{\mathrm{j}} = 0$, where j =1, 2, and 3. Since the gradients depend on both the independent variable *t* and the parameterθ, so there is no closed-form solution. Instead, the parameters are solved iteratively by successive approximation.

$$\theta^{\mathrm{k}+1} = \theta^{\mathrm{k}} + (\Delta\theta)^{\mathrm{k}} \tag{6}$$

where, $\Delta\theta = [\Delta K, \Delta A, \Delta R]^{\mathrm{T}}$ is the shift vector, k =1, 2, …, $min\{\mathrm{k}_{|\Delta\theta_{\mathrm{j}}|<\delta_{\mathrm{j}},\forall\mathrm{j}}, \mathrm{k}_{\max}\}$ where δ_{j} is the tolerance of the $|\Delta\theta_{\mathrm{j}}|$, k is the iteration number, and $\mathrm{k}_{\max}$ is the maximum iteration.

By linearization, the nonlinear model $\hat{\mathrm{C}}(\mathrm{t};\theta)$ can be approximated to linear form using a first-order Taylor polynomial expansion in the neighborhood of θ^{k} with the first-order term as follows:

$$\widehat{\ }\mathrm{C}(\mathrm{t_i};\theta) \approx \hat{\mathrm{C}}(\mathrm{t_i},\theta^{\mathrm{k}}) + \frac{\partial\hat{\mathrm{C}}(\mathrm{t_i},\theta^{\mathrm{k}})}{\partial\mathrm{K}}(\mathrm{K}^{\mathrm{k}+1} - \mathrm{K}^{\mathrm{k}}) + \frac{\partial\hat{\mathrm{C}}(\mathrm{t_i},\theta^{\mathrm{k}})}{\partial\mathrm{A}}(\mathrm{A}^{\mathrm{k}+1} - \mathrm{A}^{\mathrm{k}}) + \frac{\partial\hat{\mathrm{C}}(\mathrm{t_i},\theta^{\mathrm{k}})}{\partial\mathrm{R}}(\mathrm{R}^{\mathrm{k}+1} - \mathrm{R}^{\mathrm{k}}) \tag{7}$$

From Equation (3), $\varepsilon(\mathrm{t_i}) = \mathrm{C}(\mathrm{t_i}) - \hat{\mathrm{C}}(\mathrm{t_i};\theta^{\mathrm{k}})$ is substituted into Equation (7), rearranging to

$$\frac{\partial\hat{C}(t_i,\theta^k)}{\partial K}(\Delta K^k)+\frac{\partial\hat{C}(t_i,\theta^k)}{\partial A}(\Delta A^k)+\frac{\partial\hat{C}(t_i,\theta^k)}{\partial R}(\Delta R^k)=\varepsilon(t_i) \tag{8}$$

The *m*-linear equation system can be expressed as matrix form by.

$$J(\Delta\theta)=\varepsilon \tag{9}$$

where $J=\begin{bmatrix}\partial\hat{C}(t_1,\theta^k)/\partial K & \partial\hat{C}(t_1,\theta^k)/\partial A & \partial\hat{C}(t_1,\theta^k)/\partial R\\ \vdots & \vdots & \vdots\\ \partial\hat{C}(t_m,\theta^k)/\partial K & \partial\hat{C}(t_m,\theta^k)/\partial K & \partial\hat{C}(t_m,\theta^k)/\partial K\end{bmatrix}$ is the Jacobian matrix, and $\varepsilon=[\varepsilon(1)\varepsilon(2)\ldots\varepsilon(m)]^T$is the error vector and ^{T} denotes the vector or matrix transpose.

After a square matrix J is formed, $\Delta\theta$is solved as follows.

$$\Delta\theta=(J^TJ)^{-1}J^T\varepsilon \tag{10}$$

Therefore, the solution of Equation (4) is obtained. It is noted that the iterations fail when the matrix J^TJ becomes a singular matrix. Besides, if $\|\varepsilon^k\|$ is very large, Equation (10) is not a good approximation, and convergence is not guaranteed. Moreover, the different initial parameters may lead to different solutions, not all of which correspond to the global minimum. Furthermore, due to many multiple minimums of J, so the poor initial guesses may lead to the solution trap in a local minimum, resulting in a poor model fit.

To overcome the problem, θ is alternatively fine-tuned by the GA– an optimization and a stochastic global search technique under the principles of genetics and natural selection (John Holland, 1975). In the parameter estimation procedure using GA (Figure 1), the initial chromosomes of N_c individuals representing the solutions composed of the parameters (genes) K, A, and R, including some genetic operators, e.g., crossover rate (P_c) and mutation rate (P_m), for comprehensive optimization, are randomly generated first. Moreover, the solution obtained from the NLS method is included in the chromosome pool as a candidate to accelerate the convergence of GA. Each gene is encoded into a binary string with the total number of bits (N_{bit}), which depends on the desired accuracy at the precision level of each parameter (σ) and transformed into the universe of discourse (UOD) containing the solutions. Let the UOD of the θ is [L_K, U_K], [L_A, U_A], and [L_R, U_R], respectively, and of the genetic operator, P_c and P_m is [L_{Pc}, U_{Pc}], and [L_{Pm}, U_{Pm}], respectively, then N_{bit} is determined by.

$$N_{bit}=\sum_{i=1}^{3}\left(N_{bit,\theta_i}\right)+N_{bit,P_c}+N_{bit,P_m}=\sum_{i=1}^{3}\left(\left\lceil\log_2\left(\frac{U_{\theta_i}-L_{\theta_i}}{\sigma_{\theta_i}}\right)\right\rceil\right)+\left\lceil\log_2\left(\frac{U_{P_c}-L_{P_c}}{\sigma_{P_c}}\right)\right\rceil+\left\lceil\log_2\left(\frac{U_{P_m}-L_{P_m}}{\sigma_{P_m}}\right)\right\rceil \tag{11}$$

where N_{bit,θ_i}, N_{bit,P_c}, and $N_{bit,P_m}$are the number of bits of *θ*, *P_c*, and *P_m*, respectively, and the binary value ($B_{G,i}$) of each gene is converted to the decimal value ($D_{G,i}$) of each gene by the linear mapping of

$$D_{G,i}=L_i+\frac{(U_i-L_i)}{2^{N_{bit,i-1}}}\times B_{G,i} \tag{12}$$

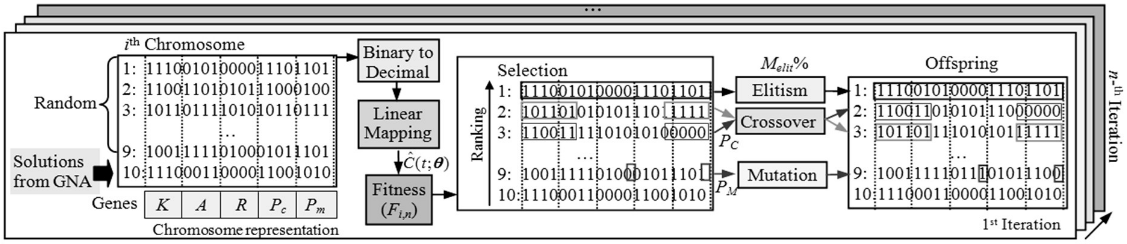

Figure 1 Block diagram of the GA assisted with GNA used for the parameter estimation of the LGR model.

The UOD is thus divided into finite pieces. The performance of the individual chromosome in the current population is evaluated using F in Equation (5) for predicting $\hat{C}(t;\theta)$, thus the chromosomes that have high fitness scores are given higher ranks. Some of them are retained M_{elit}% using the elitism strategy, while the rest are selected by the roulette wheel method to the reproduction step for generating the new offspring. Two types of reproduction in this work are crossover and mutation. In a crossover, a selected parent chromosome in the current generation exchanges genes and recombines them in the next generation to produce the new offspring, which are expected to a better performance than their parents. They change some information by crossover with P_c. The C-point on the chromosome string is chosen randomly to crossover. Besides, the mutation mutates the chromosomes with P_m. A small random change in the chromosome (e.g., converting the binary code from "0" to "1" and vice versa) providing a new offspring can help the solution to get out of the trap of local minima. The

M-point on the chromosome string to the mutation is randomly selected. Then, the earlier chromosomes are replaced by the new offspring. The process of GA is then repeated until the maximum generation (Gen_{max}) is reached.

2.2 PSO and ACO for parameters of LGR model

The particle swarm optimization (PSO) and the ant colony optimization (ACO), the other well-known population-based search techniques in the field of nature-inspired stochastic and evolutionary optimization algorithms, are applied to estimate the parameters of the LGR model, named LGR-PSO and LGR-ACO, in comparing their prediction performances with those of the proposed LGR-GA/GNA.

The PSO (1995) [21] is inspired by the movement and intelligence of the swarms in nature, such as insects, birds, and fish, in their behaviors of preying, searching food, or tracking the route to reduce total searching time. In the PSO, Figure 2A, a swarm of P particles (solutions) flies over the search space in N-dimension to locate a global optimum. All of the particles have the fitness values evaluated from the objective function (f) to be optimized and also velocities that direct their movement to follow the current best particle. Their initial position vectors $X \in R^N$ composed of the candidate parameters are randomly generated. To minimize $f(X_p(i))$ of the p^{th} particle, $p = 0, 1, \ldots, P$, at the i^{th} iteration, $f(X_p(0))$ is evaluated, such that the initial personal best position, $P_{best,p}(0) = X_p(0)$, whereas the initial global best position $G_{best}(0)$ is obtained from one having $Min(f(X_p(0)), \infty)$.

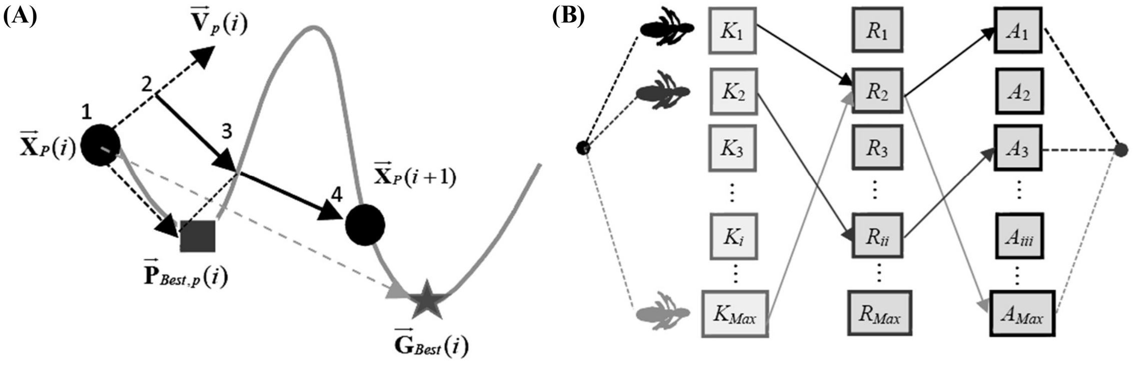

Figure 2 Block diagram of the PSO (A) and ACO with each parameter vector of K, R, and A composed of the maximum number of Max_{node} nodes (B) used for the parameter estimation of the LGR model.

In the beginning, they are moving, so the vector of velocity $V \in R^N$ is 0. During the procedure, they communicate with one another and learn to move their current positions ($X_p(i)$) toward both the personal best position ($P_{best,p}(i -1)$) obtained from the personally highest fitness value and the global best position ($G_{best}(i - 1)$) obtained from the overall highest fitness value. Therefore, the velocity is updated by.

$$V_p(i+1) = \delta^t \omega V_p(i+1) + c_1 u_1^T (P_{best,p}(i) - X_p(i)) + c_2 u_2^T (G_{best}(i) - X_p(i)) \tag{13}$$

So, the new position is updated by.

$$X_p(i+1) = X_p(i) + V_p(i+1) \tag{14}$$

where ω is the inertia weight parameter, δ is the damping ratio, c_1 and c_2 are the personal and social acceleration coefficients, and $\{u_1, u_2\} \in R^N$ are the vectors of random number uniformly distributed between 0 and 1.

In addition, the velocities are bounded by $V_{p,max}$. If $V_p > V_{p,max}$ and $V_p < -V_{p,max}$, then $V_p = V_{p,max}$ and $V_p = -V_{p,max}$, respectively. Also, the positions are bounded by $X_{p.min}$ and $X_{p,max}$. The procedure is repeated until met the maximum iteration (I). For the LGR-PSO model with 3-dimension search space of $\theta = \{K, R, A\}$, $X_p(i) = \theta_p(i) = [K_p, R_p, A_p]^T$, $P_{best,p}(i) = \theta_{pbest,p}(i) = [K_{pbest,p}, R_{pbest,p}, A_{pbest,p}]^T$, $G_{best}(i) = \theta_{gbest}(i) = [K_{gbest}, R_{gbest}, A_{gbest}]^T$, $V_{p,max} = [V_{p,max}^K, V_{p,max}^R, V_{p,max}^A]^T$, $X_{p,min} = [X_{p,min}^K, X_{p,min}^R, X_{p,min}^A]^T$, and $X_{p,max} = [X_{p,max}^K, X_{p,max}^R, X_{p,max}^A]^T$.

The fitness function of the LGR-PSO is J(θ), Equation (4). On the other hand, the ACO algorithm [22] is another nature-inspired method based on the cooperative behavior of real ant colonies for searching the shortest path of the round-trip nest to the best quality food. The pheromone used in ant communication left from the ants to the paths guides the others toward the target point, and it evaporates gradually with time. The higher level of pheromone the less travel time means those paths have a chance to be chosen. The ACO has successfully applied to the routing and load balancing problems as a graphical problem by means of finding the shortest path between points in a graph, e.g., traveling salesman problem (TSP), vehicle routing, and scheduling. It is rarely used in the parameter estimation problem.

For minimizing J(θ), Equation (4), to estimate the parameters of the LGR model (K, R, and A), the pattern of path linking the nodes (parameters) constructed by the artificial ants, which is represented as the set of solution, is considered instead of its total distance of tour, Figure 2B. In step 1, all the values of parameters K_i, R_{ii}, and A_{iii} (i, ii, iii = 1, 2, Max_{node}) divided equally within the UOD of the θ into three different vectors (Equation 15) are set as the nodes where the ants have to pass once a vector. So, there are different paths corresponding to the combinations of the number of Max_{node} nodes of each parameter vector of K, R, and A.

$$[\mathrm{KRA}] = [\mathrm{L_K L_R L_A}] + \frac{[\mathrm{U_K U_R U_A}]-[\mathrm{L_K L_R L_A}]}{\mathrm{Max_{node}}-1} \tag{15}$$

Given N_{Ant}-ant during the search, in step 2, each k^{th} ant constructs a tour from the selected i^{th}, ii^{th}, and iii^{th} nodes using the roulette wheel method, which is biased based on the deposited pheromones (phe) contained in the nodes. In the beginning, the initial pheromone is assigned to all nodes equally. In step 3, at the j^{th} iteration, the cost function J is evaluated through these nodes and used to locally update the pheromone, which can be expressed by

$$(\mathrm{phe}_{K_i}, \mathrm{phe}_{R_{ii}}, \mathrm{phe}_{A_{iii}})_j^{\mathrm{Local}} \leftarrow (\mathrm{phe}_{K_i}, \mathrm{phe}_{R_{ii}}, \mathrm{phe}_{A_{iii}})_{j-1}^{k} + \frac{\beta^+}{J_{j-1}^{k}}\phi \tag{16}$$

After completing a tour and locally updating, in step 4, the best tour (the minimum of J, J^{min}) and the worst tour (the maximum of J, J^{max}) are used to update globally the pheromone by $(\Delta\phi)^{Best}$ and $(\Delta\phi)^{Worst}$ for increasing and decreasing the chance of selecting these nodes in the next iteration, respectively, which can be expressed by

$$(\mathrm{phe}_{K_i}, \mathrm{phe}_{R_{ii}}, \mathrm{phe}_{A_{iii}})_j^{\mathrm{Global}} \leftarrow \left((\mathrm{phe}_{K_i}, \mathrm{phe}_{R_{ii}}, \mathrm{phe}_{A_{iii}})_j^{\mathrm{Local}}\right)^{\lambda} + (\Delta\phi)^{\mathrm{Best}} + (\Delta\phi)^{\mathrm{Worst}} \tag{17}$$

where

$$(\Delta\phi)^{\mathrm{Best}} = (\Delta\mathrm{phe}_{K_i}, \Delta\mathrm{phe}_{R_{ii}}, \Delta\mathrm{phe}_{A_{iii}})^{\mathrm{Best}} = \frac{\beta^-}{J^{\min}}\phi, \tag{18}$$

$$(\Delta\phi)^{\mathrm{Worst}} = (\Delta\mathrm{phe}_{K_i}, \Delta\mathrm{phe}_{R_{ii}}, \Delta\mathrm{phe}_{A_{iii}})^{\mathrm{Worst}} = \frac{\beta^-}{J_{j-1}^{\max}}\phi, \tag{19}$$

and ϕ is the weight of pheromone vector, β^+ and β^- are the positive and negative pheromone constants, respectively.

Besides, the pheromone evaporation parameter (λ) helps the ACO forget the past tour and the accumulation of the pheromones, so the ants can make a new tour means getting out the local trapped solutions. The procedure with the same ant colony is repeated until meeting the maximum number of tours (Max_{tour}).

2.3 SIR model

To evaluate the prediction performances of the proposed model, the standard SIR model [4-6], as a classical epidemiological model, is used as the baseline model. Both LGR and SIR models require less data and hence can be quickly put to use in time. The SIR model divides the population of size N into three classes, named susceptible people (S), infected people (I), and removed people (R). This basic compartmental framework, i.e.,$\mathrm{S(t)} \xrightarrow{\beta} \mathrm{I(t)} \xrightarrow{\gamma} \mathrm{R(t)}$, can be expressed by the differential equation as,

$$\begin{gathered}\frac{dS}{dt} = -\frac{\beta IS}{N}\\ \frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I\\ dR/dt = \gamma I\end{gathered} \tag{20}$$

where t is time, β is the contact rate, γ is the recovery rate, with the initial conditions of $S(0) = S_0$, $I(0) = I_0$, and $R(0) = 0$ and under the assumption of closed population, i.e., N = S(t) + I(t) + R(t).

Due to the nonlinear forms (20), the solutions cannot be solved explicitly. The analytical solution of S(t) with I(t), whereas R(t) = N – S(t) – I(t), is obtained easily by solving the first two equations of (20) as follows.

$$I(t) = R_0 \ln(S_0/S(t)) - S(t) + N \tag{21}$$

where $R_0 = \beta/\gamma$ is the reproduction number.

Alternatively, the numerical method is applied for solving (20) while the model parameters β, γ and initial value S_0 are estimated from the cumulative cases data by minimizing the objective function,

$$f(\beta, \gamma, S_0) = \left\| C_t - \hat{C}_t(\beta, \gamma, S_0) \right\|_2 \tag{22}$$

where $\hat{C}_t(\beta, \gamma, S_0) = I(\beta, \gamma, S_0) + R(\beta, \gamma, S_0)$ $\|\cdot\|_2$ is the Euclidean norm, and $I(0) = C_1$.

2.4 Data and data processing

The number of confirmed cases of COVID-19 data is retrieved from the Department of Disease Control of Thailand (DDCT), covering two waves and part of the third wave of the epidemic, from the first detected patient in Thailand on January 12, 2020, to the last recorded data on May 6, 2021. It is noted that the data provided online by WHO website did not match the official data reported by the DDCT, with time shift and wrong quantity, as shown in Figure 3. Since the predictive model is driven by data, its prediction is only as good as the data are. To obtain an accurate and reliable predictive model, the correctness of the data is crucial.

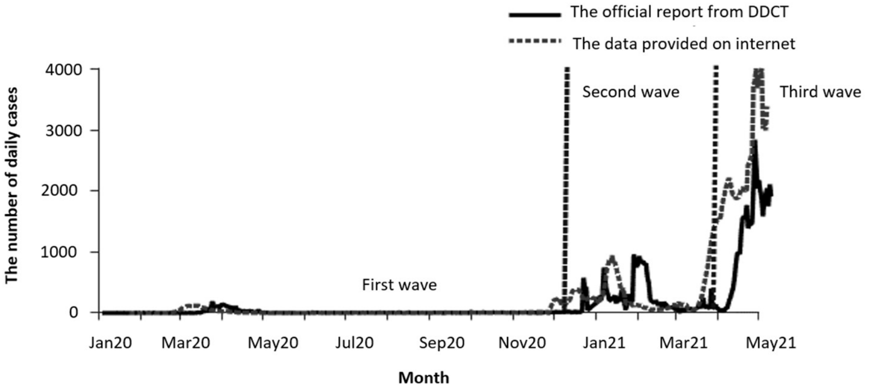

Figure 3 Incorrect data of the daily infected cases of the first-third waves of the COVID-19 epidemic in Thailand provided by WHO against the official reported data from the DDC of Thailand, leading inaccurate prediction.

The data from the first two waves of the COVID-19 epidemic are used to validate the performances of the LGR model to infer that it can be generalized as an extrapolation for predicting the trend of the third wave (launched April 1, 2021), which is ongoing at the time of writing paper. For the first and second waves, the models are trained by 102 samples (February 9 - May 20, 2020) and 111 samples (December 11, 2020 - March 31, 2021), respectively. The daily cases data are cumulatively summed and divided into two parts, i.e., training and test data, of 9 different sliding window: 10/92, 20/82, 30/82, 35/67, 40/62, 45/57, 50/65, 60/62, and 80/22 for the first wave, and 10/101, 15/196, 20/91, 25/86, 30/81, 35/76, 45/66, 60/51, and 80/31 for the second wave. For modeling the LGR models during the ongoing third waves (April 1, 2021 - May 6, 2021) at both national and provincial levels, such as Bangkok (BKK, the capital), Chiang Mai (CM, the tourism city in the north), Chonburi (CBI, the tourism city in the east) and Samut Prakan (SPK, where a large number of migrant workers live), the sliding windows of training/test data are 10/26, 20/16, and 30/6.

2.5 Criteria of the goodness-of-fit

Due to the assumption of normal distribution of the error, the goodness-of-fit can be evaluated using the standard metrics, such as the coefficient of determination (R^2), and the root mean square error (RMSE),

$$R^2 = \frac{\sum_t \left(\hat{C}(t;\theta)\right)^2}{\sum_t (C(t))^2} \tag{23}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_W}\sum_{t=1}^{N_W}\left(C(t) - \hat{C}(t;\theta)\right)^2} \qquad (24)$$

For the significance of the fitted parameters of the LGR model, the t-statistic and its p-value are used to make a conclusion in a significance test at the 95% confidence interval (CI) or significance threshold (α) of 0.05.

3. Results and discussion

The results of parameter estimation and the comparison of prediction performances between the proposed LGR-GA/GNA, the LGR-PSO, and the LGR-ACO for the first and second waves of the COVID-19 epidemic in Thailand at provincial and national levels are presented in this section. Additionally, the prediction performances between the proposed model and the SIR model for those waves and the ongoing third wave of the COVID-19 epidemic are examined.

The predictive LGR models with the parameter estimation methods (i.e., GNA and GA, including PSO and ACO) are implemented using our writing source code run on the Matlab software package. To solve the differential equations in (20) and estimating the parameters in (22) of the SIR model, Matlab's built-in ode45 and fminsearch functions are used, respectively.

The selected parameters and their setting values of the GNA are as follows, the maximum iteration (k_{max}) = 20 and the tolerance of $\delta_{|\Delta K|}$, $\delta_{|\Delta A|}$, and $\delta_{|\Delta R|}$ = 200, 100, and 0.02, respectively, and those of the GA are given in Table 2. Whereas the two major genetic operators (P_c and P_m) are fine-tuned simultaneously with the LGR's parameters (K, R, and A) to improve the performance of the GA. There are various types of crossover and mutation, but no theoretical proof to guarantee which one of them is the best for different optimization problems. However, the one-point, multiple-point, and uniform crossovers and mutations are frequently used and simple methods. In this work, through the base run on different test data by controlling the other GA parameters, the two-point crossover (i.e., the random segment of information of selected parents' chromosomes is exchanged to produce two unique offspring individuals), and the one-point mutation (i.e., flip the binary bit at the random position in the selected chromosomes) are the best combinations providing averagely high fitness value among others. The parameters of the PSO and ACO, including their selected parameters obtained from the factorial design, are shown in Table 3.

The estimated parameters, the goodness-of-fit performances, and significance tests of the LGR-GA/GNA model for the first and second waves of the COVID-19 epidemic are shown in Table 4, while the predictions of the basic measures are shown in Table 5.

Table 2 GA parameters and relevant parameters with their descriptions and setting values for fine-tuning the LGR's parameters (K, R, and A), simultaneously with some GA's parameters themselves, i.e., P_c and P_m.

Parameters	Descriptions and setting values	Fine-tuning interval
1) The number of chromosomes (N_c+1) and genes	10 and 5	-
2) The total number of binary bit (N_{bit}) ($N_{bit,K}$+ $N_{bit,A}$ + $N_{bit,R}$ + $N_{bit,Pc}$ + $N_{bit,Pm}$)	44 (11+10+7+7+9)	-
3) The number of elitism chromosomes (M_{elit}%)	10%	-
4) The precision level (σ_K, σ_A, σ_R, σ_{Pc}, and σ_{Pm})	100, 50, 0.01, 0.005, and 0.001	-
5) Maximum generation (Gen_{max})	1000	-
6) The UOD: [L_K, U_K], [L_A, U_A], [L_R, U_R], [L_{Pc}, U_{Pc}], and [L_{Pm}, U_{Pm}]	[1000, 200000], [10, 40000], [0.001, 1], [0.1, 0.9], and [0.001, 0.5]	-
7) Type of crossover and crossover rate (P_C)	Two-point crossover	0.1-0.9
8) Type of mutation and mutation rate (P_m)	One-point mutation	0.001-0.5

Table 3 The parameters and their setting values of the PSO and ACO for estimating the LGR's parameters.

PSO		ACO	
Parameters	Setting values	Parameters	Setting values
1) The number of particles (P)	20	1) The number of artificial ants (N_{Ant})	20
2) Inertia weight (ω)	1.0	2) Weighted pheromone (phe)	0.06
3) Damping ratio (δ)	0.99	3) Pheromone evaporation parameter (λ)	0.95
4) Personal and social acceleration coefficients (c_1 and c_2)	2	4) Positive and negative pheromone constants (β^+ and β^-)	0.2 & 0.3
5) The number of maximum iterations (I)	2000	5) The maximum number of tours (Max_{tour})	2000

Table 4 The results of the estimated parameters (*K*, *R*, and *A*) of the LGR model using the GA/GNA method for the 9-sliding windows of training/test samples and the performances of the fitted model (RMSE and R^2) for the first and second waves of the COVID-19 epidemic in Thailand at the national level.

Phase	Training/ Test samples	Parameters (95% of CI)			t-stat (*p*-value)	RMSE		R^2	
		K (Final size)	R (Infected rate)	A		Train	Test	Train	Test
First wave (Feb 9,2020 - May 20,2020)	10/92	3,609.45 (±979)	0.33 (±0.15)	11,862.14 (±2,694)	<0.05	1.03	1,113.17	0.88	0.78
	20/82	3,233.62 (±419)	0.17 (±0.077)	4,209.33 (±1,800)	<0.05	2.06	468.60	0.89	0.90
	30/72	3,073.02 (±208)	0.27 (±0.026)	3,4901.11 (±1,408)	<0.05	5.16	313.86	1.00	0.93
	35/67	3,141.94 (±127)	0.28 (±0.036)	3,6115.42 (±940)	<0.05	18.40	388.29	0.99	0.88
	40/62	3,069.49 (±74)	0.21 (±0.030)	4,370.64 (±578)	<0.05	43.22	224.19	0.99	0.92
	45/57	3,063.17 (±64)	0.19 (±0.042)	1,914.11 (±376)	<0.05	54.33	176.98	0.99	0.93
	50/52	3,012.17 (±80)	0.17 (±0.022)	1,061.29 (±70)	<0.05	70.94	113.99	0.99	0.94
	60/42	3,004.48 (±63)	0.16 (±0.018)	681.68 (±42)	<0.05	90.78	65.68	0.99	0.98
	80/22	3,003.69 (±53)	0.17 (±0.019)	852.65 (±21)	<0.05	90.99	47.40	1.00	0.90
Second wave (Dec 11,2020 -Mar 31,2021)	10/101	19,789.03 (±1,421)	0.24 (±0.14)	143.77 (±108)	<0.05	309.97	6,070.73	0.89	0.50
	15/96	6,016.68 (±888)	0.27 (±0.16)	81.93 (±53)	<0.05	294.62	12,740.48	0.85	0.35
	20/91	16,237.27 (±942)	0.13 (±0.09)	59.38 (±33.3)	<0.05	323.95	5,043.23	0.90	0.80
	25/86	16,471.60 (±999)	0.12 (±0.05)	42.28 (±36)	<0.05	337.25	4,963.53	0.97	0.86
	30/81	20,213.57 (±887)	0.10 (±0.012)	42.45 (±22.1)	<0.05	379.53	2,498.59	0.98	0.95
	35/76	8,768.25 (±638)	0.14 (±0.022)	27.25 (±7.61)	<0.05	317.96	11,792.06	0.99	0.78
	45/66	21,515.06 (±475)	0.07 (±0.014)	24.51 (±6.26)	<0.05	703.69	2,332.02	0.96	0.93
	60/51	33,846.30 (±334)	0.07 (±0.010)	42.63 (±6.14)	<0.05	764.55	7,386.76	0.98	0.86
	80/31	23,554.64 (±113)	0.09 (±0.010)	40.91 (±3.7)	<0.05	911.52	773.03	0.99	0.83

In Table 3, it can be seen that there are large confidence intervals of the parameters at the early stage of the first wave of the COVID-19 epidemic means less knowledge about the context of infection, causing high uncertainty in the prediction. In this infection wave, three weeks of data after the outbreak is sufficient for the proposed model to gain a more accurate prediction. Moreover, the high RMSE and low R^2 criteria on the test data show poor performance in the early phase due to the difference between the portions of curves in the training that are smooth and those in the test that are sharp with high nonlinearity (Figure 4). Moreover, the R^2 criterion of the test data seems to be inconsistent, reflecting the real predictions. In the significance tests for the parameters of the LGR models, the p-values are all less than 0.05, demonstrating an acceptable prediction. For the second wave of the COVID-19 epidemic, the confidence intervals of the parameters, including the error measures, fluctuate due to several spikes in daily confirmed cases data (Figure 3) that appeared to be different and not consistent. The spikes in infection caused by the clusters of migrant workers in the overcrowded communities and abrupt restriction and quarantine measures limit the performance of the proposed LGR models.

In Table 4, for the first wave, it can be seen that the dates of occurrence of the maximum cases, the cessation date, the maximum number of cases, and the number of final sizes obtained from the proposed LGR models vary from 18-21 March 2020, from 18-27 April 2020, between 130-200, and between 3,000-3,100, respectively, which are close to the observed data. In contrast, those of the second wave with several spikes have a wide variation and are not close to the observed data. It should be mentioned that the proposed model cannot well characterize the growth function of the COVID-19 epidemic with the infection spikes. However, it is expected that there are no such spikes in the third wave of infection as less stringent measures are taken at this stage. Therefore, from the prediction results with less RMSE and high R^2 of the first wave of the epidemic, it can be inferred that the proposed LGR-GA/GNA models can be used to predict the trend in the third wave of the COVID-19 epidemic.

The prediction results of the number of cumulative cases for the first and second waves of the COVID-19 epidemic between the LGR-GNA and LGR-GA/GNA models are compared with the observed data (Figure 4) for all nine sliding windows of training/test data. It is seen that the proposed LGR-GA/GNA tracks the trends and provides a more accurate final size of infected case than the other. However, in the second wave with multiple infection spikes, both models underestimate the final size of the epidemic. In addition, the prediction results of the daily cases (Figure 5), which are obtained from Equation (1), demonstrated that the LGR models suit well in tracking the trends of the epidemic with one spike in the infection cycle.

For the third wave of the COVID-19 epidemic, the parameter estimation results, the error measures based on RMSE and R^2 criteria, and the prediction results of the basic measures (the date when maximum cases occurred, the maximum number of cases, the cessation date, and the final size, parameter *K*) obtained from the proposed LGR-GA/GNA models for the four most vulnerable provinces (BKK, CM, CBI, and SPK) and the whole country are shown in Table 6. The results of in-sample and out-of-sample predictions for each sliding window are shown in Figure 6. The additional results of the prediction trends of the daily cases of infection are shown in Figure 7. The prediction results of the total number of cases with the estimated cessation date of all three waves of the epidemic in Thailand at the national level using the last training sliding window are summarized in Table 7.

Table 5 The prediction results of the basic measures obtained from the proposed LGR-GA/GNA models for all nine sliding windows of training/test compared with the actual data for the first and second waves of the COVID-19 epidemic in Thailand at the national level.

Date	Training/Test samples (Window size)	Prediction results			
		Date at maximum cases (ln$A$)/r	Maximum daily cases (K/2)	Ending date	Final size
First wave (Feb 9,2020 -May 20, 2020)	10 (Feb 9-19)/92	Mar 8	296	Apr 20	3,609.45 (±8,979)
	20 (Feb 9-29)/82	Mar 29	137	May 3	3,233.62 (±4,119)
	30 (Feb 9-Mar 10)/72	Mar 18	205	Apr 16	3,073.02 (±2,098)
	35 (Feb 9-Mar 15)/67	Mar 17	219	Apr 11	3,141.94 (±1,427)
	40 (Feb 9-Mar 20)/62	Mar 20	161	Apr 18	3,069.49 (±704)
	45 (Feb 9-Mar 25)/57	Mar 20	143	Apr 20	3,063.17 (±640)
	50 (Feb 9-Mar 30)/52	Mar 21	136	Apr 25	3,012.17 (±280)
	60 (Feb 9-Apr 9)/42	Mar 21	120	Apr 27	3,004.48 (±123)
	80 (Feb 9-Apr 29)/22	Mar 20	127	Apr 26	3,003.69 (±53)
Actual observations		Mar 22	188	May 4	2,966
Second wave (Dec 11, 2020 – Mar 31, 2021)	10 (Dec 11-21)/101	Jan 1	1186	Jan 28	19,789.03 (±10,421)
	15 (Dec 11-26)/96	Dec 27	405	Jan 19	6,016.68 (±6,888)
	20 (Dec 11-31)/91	Jan 11	527	Feb 28	16,237.27 (±2,742)
	25 (Dec 11-Jan 5)/86	Jan 12	493	Feb 27	16,471.60 (±1,999)
	30 (Dec 11-Jan 10)/81	Jan 17	505	Mar 21	20,213.57 (±1,287)
	35 (Dec 11-Jan 15)/76	Jan 4	306	Feb 13	8,768.25 (±1,238)
	45 (Dec 11-Jan 25)/66	Jan 26	376	Apr 5	21,515.06 (±725)
	60 (Dec 11-Feb 9)/51	Feb 2	592	Mar 29	33,846.30 (±334)
	80 (Dec 11-Mar1)/31	Jan 22	530	Mar 26	23,554.64 (±113)
Actual observations		Jan 25	959	Mar 31	25,000

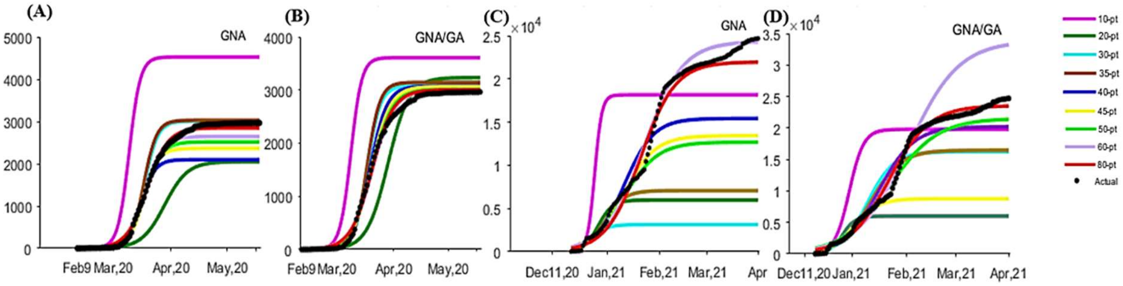

Figure 4 Comparison the prediction results including the final size of epidemic between the LGR-GNA and LGR-GNA/GA models against the observed data (black dotted-line) for all nine sliding windows of training/test samples of (A-B) first and (C-D) second waves of the COVID-19 epidemic in Thailand at the national levels.

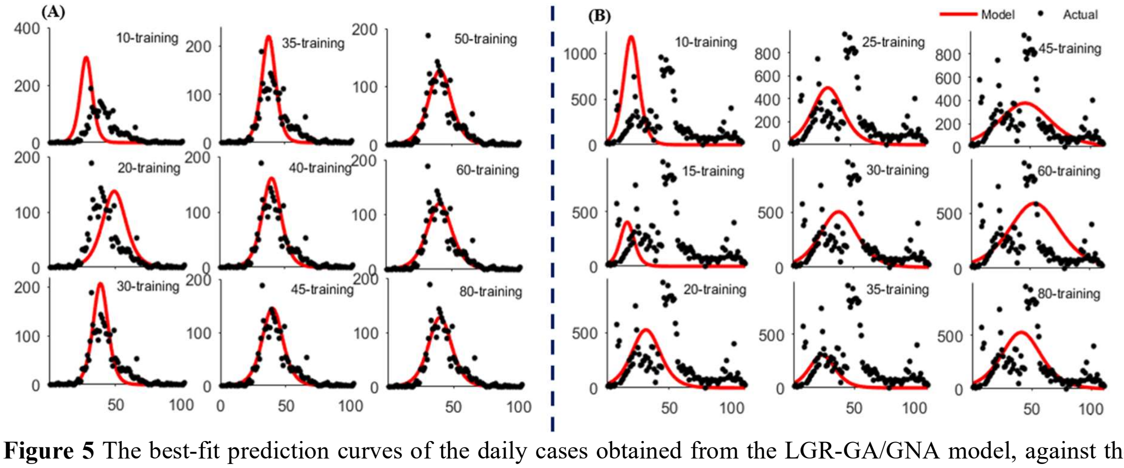

Figure 5 The best-fit prediction curves of the daily cases obtained from the LGR-GA/GNA model, against the observed data for all nine sliding windows of training/test of (A) first (Feb 9- May 20, 2020) and (B) second (Dec 11- Mar 31, 2021) waves of the COVID-19 epidemic in Thailand at the national level.

Table 6 The results of the estimated parameters (K, R, and A) of the proposed LGR-GA/GNA model and the performances of the fitted model (RMSE and R^2), including the prediction results for the 3-sliding windows of training samples during the ongoing third wave of the COVID-19 epidemic in Thailand both the provincial and national levels.

Level	Training	Parameters			RMSE		R^2		Prediction results		
Provincial	samples (date)	K	R	A	Train	Test	Train	Test	Date at Max. cases	Max.daily cases	Ending date
BKK	Apr,1-10	33,030	0.352	917.95	61.45	23,894	0.975	0.464	Apr,19	2,894	Apr,28
	Apr,1-20	20,873	0.158	94.08	198.50	10,141	0.983	0.031	Apr,29	824	May,26
	Apr,1-30	34,833	0.143	125.92	349.70	16,585	0.992	0.112	May,4	1,245	Jun,28
CM	Apr,1-10	3,462	0.990	384,230	10.60	816.4	0.976	0.284	Apr,13	857	Apr,22
	Apr,1-20	3,539	0.297	148.41	155.10	150.4	0.980	0.506	Apr,17	263	May,11
	Apr,1-30	4,005	0.208	52.45	255.50	169.8	0.979	0.798	Apr,19	208	May,28
CBI	Apr,1-10	7,919	0.567	8563.40	9.39	6,630.5	0.986	0.020	Apr,16	1,122	Apr,14
	Apr,1-20	6,103	0.192	132.87	96.81	3,856	0.963	0.020	Apr,25	292	May,10
	Apr,1-30	5,314	0.128	47.24	164.70	3,133.6	0.958	0.082	Apr,30	170	Jun, 23
SPK	Apr,1-10	19,685	0.437	7,148.60	4.785	14,308	0.996	0.596	Apr,20	2,141	Apr,24
	Apr,1-20	13,770	0.142	276.49	59.84	2,774.9	0.948	0.723	May,10	488	Jun,8
	Apr,1-30	10,219	0.107	125.14	69.00	2,192.9	0.982	0.168	May,15	275	Jun, 28
National	Apr,1-10	108,970	0.350	1,241.30	62.13	76,333	0.998	0.531	Apr,20	9,502	May,2
	Apr,1-20	157,930	0.178	270.45	809.49	66,243	0.985	0.043	May,1	7,016	Jun,6
	Apr,1-30	158,760	0.113	90.65	2,001	841.0	0.986	0.134	May,10	4,209	Jul,15

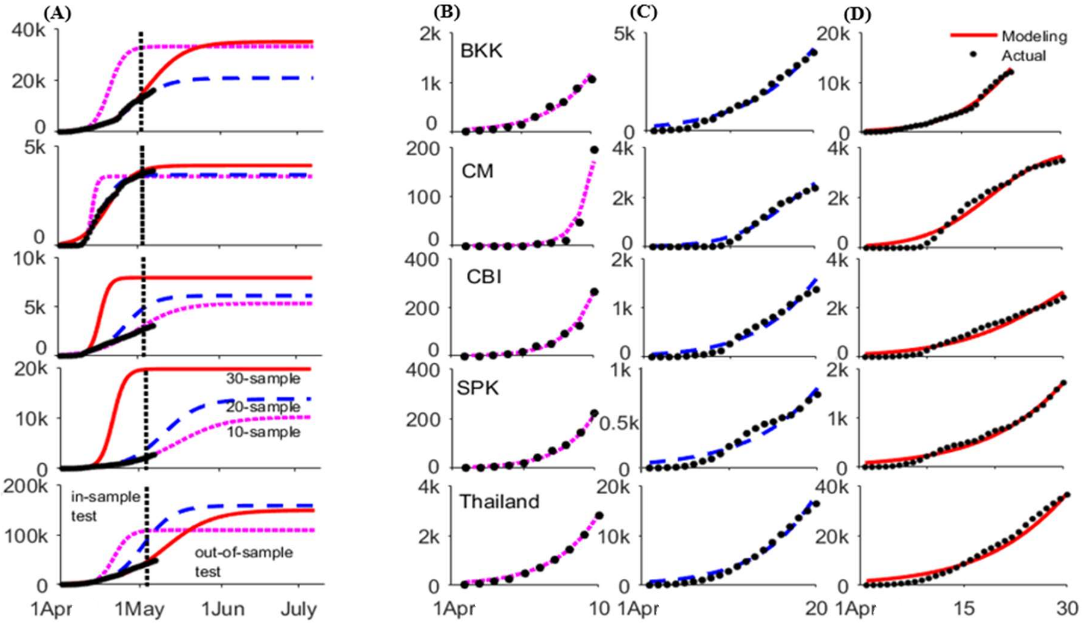

Figure 6 The output of best-fit curves of the cumulative infected cases from the proposed LGR-GA/GNA model for (A) the in-sample and out-of-sample tests using the (B-D) 10,20 and 30-sliding windows of training samples for the ongoing third wave of the COVID-19 epidemic in Thailand both provincial and national levels.

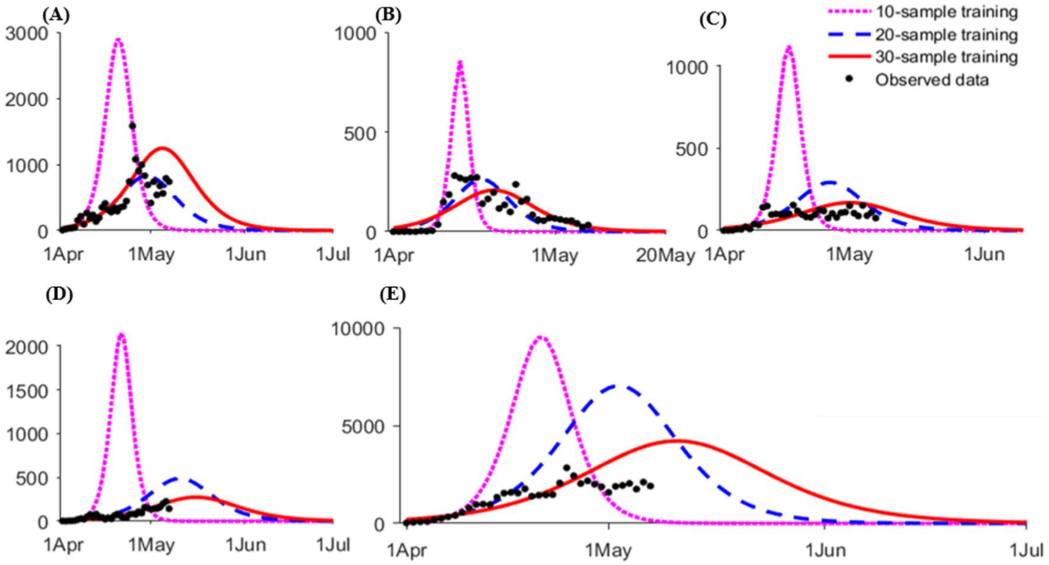

Figure 7 The output of best-fit curves of the daily infected cases from the proposed LGR-GA/GNA model for the in-sample and out-of-sample tests using the 3-sliding windows of training samples for the ongoing third wave of the COVID-19 epidemic in Thailand both provincials; (A) Bangkok, (B) Chiang Mai, (C) Chonburi, (D) Samut Prakan, and (E) national levels.

Table 7 Prediction results of the cessation date and final size of the third wave of the COVID-19 epidemic using the proposed LGR-GA/GNA model in Thailand at the national level.

Measures	Firstwave Jan 1 – Nov 30, 2020	Second wave Dec 1 – Mar 31, 2020	Third wave From1 Apr 2021	Total
Estimated final size	3,003	23,554	158,760	185,317
Actual fina size	3,912	24,814	-	
Estimated cessation date	April 26, 2020	March 26, 2021	July 15, 2021	
Actual cessation date	May 1, 2020	April 1, 2021	-	

*The predictions have been established on May 6, 2021.

This study found that the outbreak growth rates differed among four provinces in Thailand. In the earliest epidemic phase, the model does not fit well with the observed data and the number of cumulative cases is significantly underestimated. However, the goodness-of-fit of the model increases as the sample size for training the model increases. Except in the case of BKK, a linear upward trend causes the model to underestimate the final size (Figure 6). According to the prediction results, the outbreak will possibly end in late June to mid-July 2021, except in CM, with an end in May. The final size of the epidemic during the third wave of infection is about 35,000 cases for BKK, 4,000 cases for CM, 5,500 for CBI, 10,300 for SPK, and 160,000 cases for the whole country.

Moreover, in the comparison of parameter estimation between the GA, PSO, and ACO for the LGR models, the results of the estimated parameters, including the goodness of fits, of the LGR-PSO and LGR-ACO are shown in Table 8 for all three waves of the COVID-19 epidemic in Thailand at the national level, whereas those of the LGR-GA are tabulated in Table 9. The prediction results of the trends for all three waves of the COVID-19 epidemic are illustrated in Figure 8. The prediction results from the first two waves demonstrate the superiority of GA over the others (with lower RMSE and higher R^2, including closer prediction to the final size of the epidemic).

To evaluate prediction performances of the proposed LGR-GA/GNA model with the SIR model, the sliding windows of training data are 10, 20, 30, and 40 for the first wave, 15, 30, 45, and 60 for the second wave, and 15 and 30 for the third wave. The estimated parameters of the SIR model and the comparison of the performances between LGR and SIR models are presented in Table 9. The validation results of the number of cumulative cases of COVID-19 for the first and second waves and the prediction results for the ongoing third wave are shown in Figure 9. It is seen that, at the beginning of the epidemic, the SIR does not work well, indicated by the high RMSE and low R^2 when compared to the proposed LGR. For the high spread of infections with multiple peaks (especially in the second wave), the proposed LGR model has advantages over the rest with lower RMSE and higher R^2.

Table 8 Parameter estimation (K, R, and A) for the LGR-PSO and LGR-ACO models and the goodness-of-fit measures (RMSE and R^2) between the proposed LGR-GA/GNA model and the SIR model for the past two waves and during the ongoing third wave of the COVID-19 epidemic in Thailand at the national levels using the 4-sliding and 2-sliding windows of the training samples, respectively.

Epidemic wave	No. of data training	LGR-PSO								LGR-ACO							
		Final size	K	R	A	Training		Test		Final size	K	R	A	Training		Test	
						RMSE	R^2	RMSE	R^2					RMSE	R^2	RMSE	R^2
1	10	1000	2.66	13.30	1000	6	0.99	1272	-0.40	6121	4.13	24.50	6124	21	0.91	2302	-3.58
	20	1511	2.86	14.96	1511	23	1.00	879	0.33	2883	9.64	28.68	3036	228	0.73	395	0.87
	30	2415	4.26	18.88	2415	45	1.00	257	0.94	1972	5.41	16.26	1973	228	0.92	575	0.71
	40	2642	4.81	20.00	2643	53	1.00	130	0.99	3039	9.31	24.11	3128	233	0.94	199	0.97
2	15	1920	1.20	6.99	1,920	134	0.97	11740	-1.31	15051	8.80	31.51	15147	400	0.76	3421	0.80
	30	8581	6.31	21.36	8583	276	0.98	7079	0.16	6370	8.55	17.35	6376	826	0.86	8585	-0.24
	45	11791	8.79	27.33	11,838	456	0.98	5212	0.54	9104	7.35	21.12	9109	785	0.95	6767	0.23
	60	21956	10.00	40.93	22614	1087	0.97	969	0.98	9816	2.33	11.30	9816	4986	0.38	6800	0.22
3	15	14105	2.72	13.90	14105	96	1.00	71685	-0.85	53478	18.86	85.76	158700	2127	0.37	59989	-0.29
	30	4,209	5.17	23.07	45212	683	1.00	51072	0.06	60028	9.17	30.04	60583	3331	0.92	43466	0.32

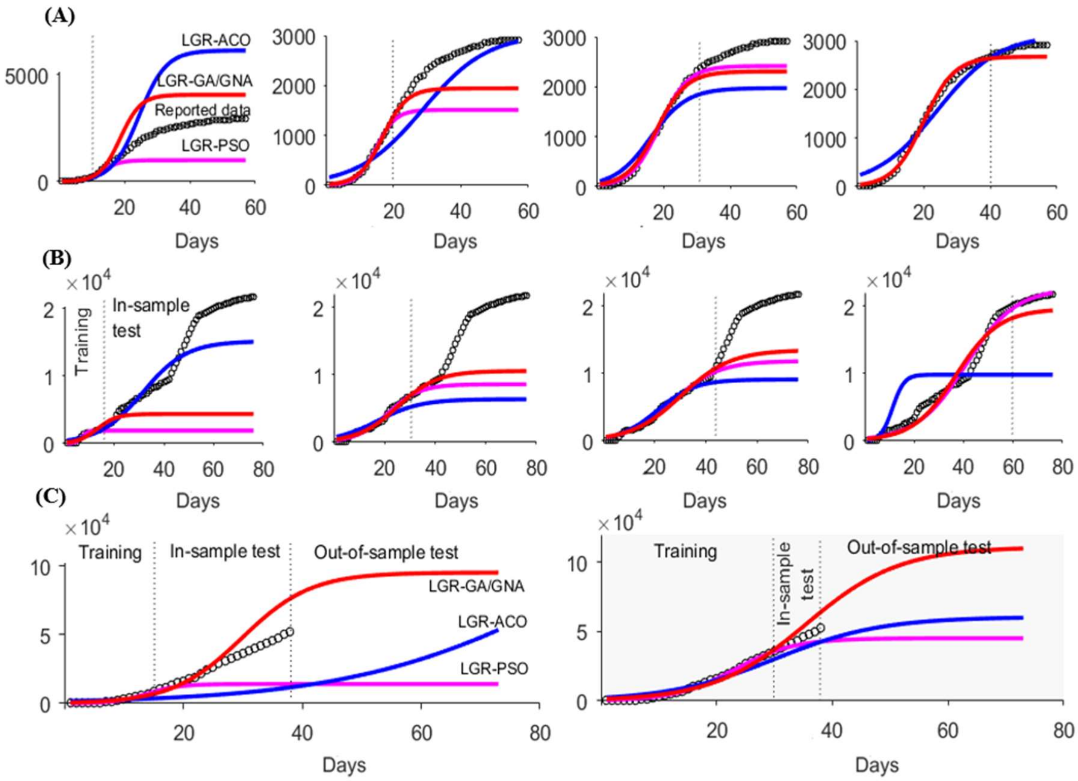

Figure 8 Comparison on the projections of the cumulative cases and the estimation of final size of the COVID-19 epidemic in Thailand at the national level between the proposed LGR-GA/GNA model, the LGR-PSO, and the LGR-ACO model for the 4-sliding and 2-sliding windows of the training samples of the past first (A) and (B) second waves and (C) during the ongoing third wave, respectively.

Table 9 Parameter estimation for the SIR model (S_0, β, and γ) including the reproductive number (R_0) and the performances comparison of the goodness of fit (RMSE and R^2) between the proposed LGR-GA/GNA model and the SIR model for the past two waves and during the ongoing third wave of the COVID-19 epidemic in Thailand at the national levels using the 4-sliding and 2-sliding windows of the training samples, respectively.

Epidemic	No. of	SIR model								The proposed LGR-GA/GNA			
wave	data	Parameters			R_0	Performances				Performances			
	training	S_0	β	γ		RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
						Training	Training	Test	Test	Training	Training	Test	Test
1	10	6.3×10^7	374.7	388.5	0.96	65.91	0.12	2018	-2.52	7.9	0.98	1143.87	-0.13
	20	3484	1.38	1.04	1.33	22.31	0.99	822.7	0.41	37.0	0.99	569.76	0.71
	30	20900	3.53	3.32	1.06	36.74	0.99	212.2	0.96	73.4	0.99	329.67	0.90
	40	31844	4.43	4.24	1.04	39.78	0.99	109.2	0.98	61.0	0.99	119.90	0.98
2	15	2,118	1.27	0.42	2.99	122.2	0.97	11680	-1.29	313.3	0.85	9917.32	-0.65
	30	709154	10.98	10.90	1.00	253.4	0.98	5268	0.53	328.0	0.97	5907.68	0.41
	45	2,251,787	14.88	14.84	1.00	362.3	0.98	3663	0.77	494.4	0.98	4406.63	0.67
	60	2.6×10^7	15.05	15.01	1.00	698.7	0.98	3645	0.77	1358.0	0.95	1540.61	0.96
3	15	60610	2.23	1.90	1.17	71.2	0.99	69644	-0.74	831.8	0.90	25842.4	0.75
	30	898664	4.07	3.91	1.04	2583	0.95	36022	0.53	1817.2	0.97	18450.6	0.87

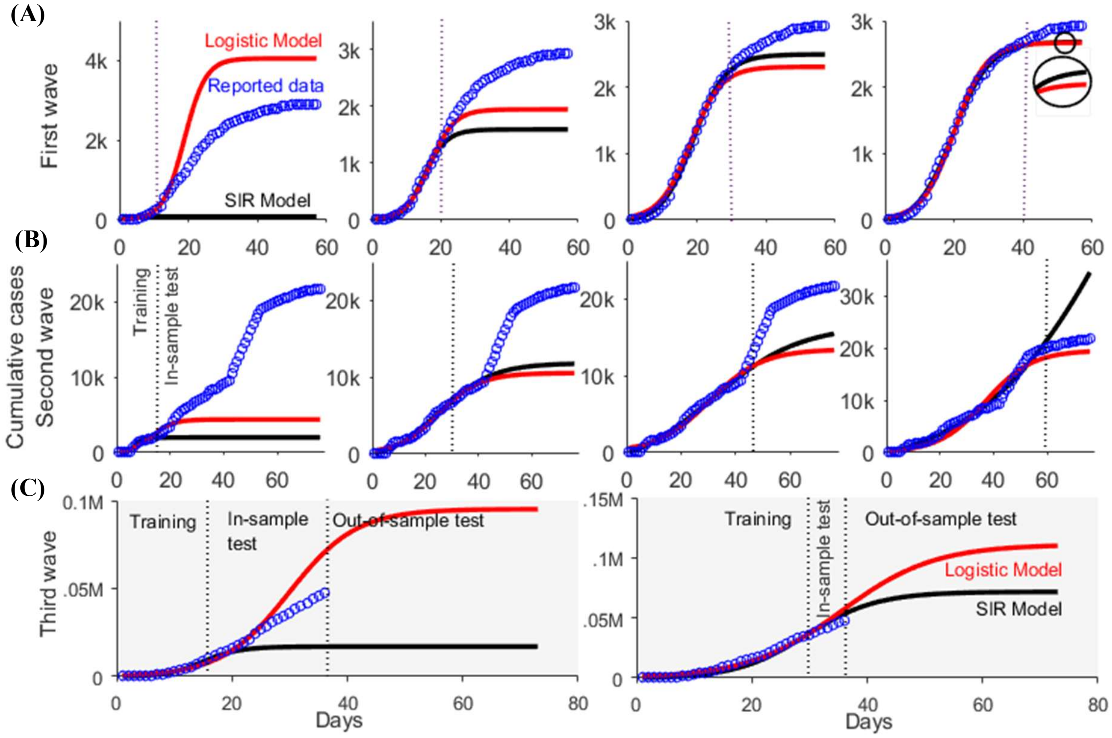

Figure 9 Comparison on the projections of the cumulative cases and the estimation of final size of the COVID-19 epidemic in Thailand at the national level between the proposed LGR-GA/GNA model and the SIR model for 4 and 2 sliding training windows of (A) the past first and (B) second waves and (C) the ongoing third wave, respectively.

4. Conclusion

In this study, the LGR-based predictive models are used to estimate the basic measures and predict the trends of the COVID-19 epidemic during the ongoing third wave of infection in Thailand at both provincial and national levels. The LGR models are optimized using the GA assisted by the GNA for enhancing the prediction performance. The proposed LGR-GA/GNA models provide goodness-of-fit with low RMSE and high R^2 for the first two waves that have passed the peak of infection, inferring that they can be used as extrapolation for the current wave. The comparison of the prediction performance between the proposed LGR-GA/GNA, the LGR-PSO, and the LGR-ACO showed that the LGR-GA outperforms the rest. Furthermore, when compared to the SIR model to verify the effectiveness of the LGR model in its prediction, the proposed LGR-GA can improve the goodness-of-fit better than the other. However, due to the small amount of data in the early phase of the epidemic used to train the model, the out-of-sample prediction results may be inconsistent, although they fit well with the in-sample test. In addition, the lack of including other external factors into the model may miss achieving the optimal prediction model. In the future, alternatively, a machine learning approach using a small amount of available training data in the early period of an outbreak can be applied to the prediction models for new incoming epidemics.

5. References

[1] Marioli AF, Bullano F, Kucinskas S, Moreno RC. Tracking $\mathcal{R}$ of COVID-19: a new real-time estimation using the Kalman filter. PLoS ONE. 2021;16(1):1-16.

[2] Tantrakarnapa K, Bhophornangkul B. Challenging the spread of COVID-19 in Thailand. One Health. 2020;1:1-10.

[3] Mahikul W, Chotsiri P, Ploddi K, Pan-ngum W. Evaluating the impact of intervention strategies on the first wave and predicting the second wave of COVID-19 in Thailand: a mathematical modeling study. Biology. 2021;10(2):1-12.

[4] Zhang L, Huang J, Yu H, Liu X, Wei Y, Lian X, et al. Optimal parameterization of COVID-19 epidemic models. AOSL. 2021;14(4):1-23.

[5] Moein S, Nickaeen N, Roointan A, Borhani N, Heidary Z, Javanmard SH, et al. Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan. Sci Rep. 2021;11(4725):1-9.

[6] Chu J. A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain. PLoS ONE. 2021; 16(3):1-36.

[7] Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. 2020;12(3):165-174.

[8] Sen D, Sen D. Use of a modified SIRD model to analyze COVID-19 data. Ind Eng Chem Res. 2021;60(11):4251-4260.

[9] Rath S, Tripathy A, Tripathy AR. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes Metab Syndr. 2020;14(5):1467-1474.

[10] Girardi P, Greco L, Mameli V, Musio M, Racugno W, Ruli E, et al. Robust inference for non-linear regression models from the Tsallis score: application to coronavirus disease 2019 contagion in Italy. Stat. 2020;9(1):1-9.

[11] Yang Q, Wang J, Ma H, Wang X. Research on COVID-19 based on ARIMA modelΔ-Taking Hubei, China as an example to see the epidemic in Italy. J Infect Public Health. 2020;13(10):1415-1418.

[12] Wieczorek M, Siłka J, Połap D, Woźniak M, Damaševičius R. Real-time neural network based predictor for cov19 virus spread. PLoS ONE. 2020;15(12):1-18.

[13] Wongsathan R. Real-time prediction of the COVID-19 epidemic in Thailand using simple model-free method and time series regression model. Walailak J Sci & Tech. 2021;18(14):1-11.

[14] Naudé W. Artificial intelligence vs COVID-19: limitations, constraints and pitfalls. AI Soc. 2020;35(3):761-765.

[15] Shen CY. Logistic growth modeling of COVID-19 proliferation in China and its international implication. Int J Infectious Disease. 2020;96:582-589.

[16] Ahmadi A, Fadai Y, Shirani M, Rahmani F. Modeling and forecasting trend of covid-19 epidemic in Iran until May13, 2020. Med J Islam Repub Iran. 2020;34(1):183-195.

[17] Gutiérrez CRA, Colorado D, Bautista HSL. Comparison of an artificial neural network and Gompertz model for predicting the dynamics of deaths from COVID-19 in México. Nonlinear Dyn. 2021;104(4):4655-4669.

[18] Zhan C, Zheng Y, Lai Z, Hao T, Li B. Identifying epidemic spreading dynamics of COVID-19 by pseudocoevolutionary simulated annealing optimizers. Neural Comput Appl. 2021;33(10):4915-4928.

[19] Paggi M. An analysis of the Italian lockdown in retrospective using particle swarm optimization in machine learning applied to an epidemiological model. Physics. 2020;2(3):368-382.

[20] He S, Peng Y, Sun K. SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dyn. 2020;101(3):1667-1680.

[21] Viboud C, Simonsen L, Chowell G. A generalized growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics. 2016;15:27-37.

[22] Roseline OO, Adewale FL, Golam BMK, Joseph BA, Benedita BA. Predictive modelling of COVID-19 confirmed cases in Nigeria. Infect Dis Model. 2020;5:543-548.