RESEARCH ARTICLE

A NONLINEAR REGRESSION MODEL, ANALYSIS AND SIMULATIONS FOR THE SECOND WAVE OF COVID-19: THE CASE STUDY OF TURKEY

Sıla Övgü KORKUT UYSAL [1,*] (iD) , **Nurcan GÜCÜYENEN KAYMAK** [2] (iD) , **Yeşim ÇİÇEK** [1] (iD)

[1] Engineering Sciences, Faculty of Architecture and Engineering, Izmir Katip Çelebi University, Izmir, Turkey
[2] Management Information Systems, Faculty of Economics and Administrative Sciences, Dogus University, Istanbul, Turkey

ABSTRACT

COVID-19 pandemic disease gained major attention among scientists due to its high mortality/ infectiousness rate. Moreover, the analysis of this disease requires much attention by the Government to take precautions and construct strategies. This study aims to develop a new nonlinear model for COVID-19. The main focus is the time when the number of daily infected individuals has begun to increase constantly. To this end, the time series from 1 August 2020 to 22 September 2020 is conducted. Moreover, the proposed model takes into account the disease characteristics. After the model parameters are obtained by detailed mathematical analysis by the trained data, the model is validated by the test/evaluation data set. The results and simulations show that the proposed model has a perfect match with the raw data. Furthermore, the calculated standard errors when compared by the population of Turkey are evidence of how well the model fits the raw data. This study is important not only because it achieves good results but also because it is the first nonlinear regression model including its mathematical analysis for the COVID-19 pandemic.

Keywords: COVID-19, Nonlinear Regression, Nonlinear Analysis, Simulation, Standard Error

1. INTRODUCTION

A mathematical model of a biological system or a complex disease is a powerful tool to decide under which conditions it will continue or wiped out. Undoubtedly that each disease has its own particular characteristics thus some assumptions are provided to address real-life situations.

Coronavirus disease 2019 (COVID-19) is an infectious disease emerging in China in December 2019 [1]. It has been declared a pandemic by the World Health Organization (WHO) on March 11 [2]. Today, COVID-19 has over 31 million confirmed cases and 3.14 % deaths as of September 24, 2020 [3]. On the other hand, to that date, it is reported by the Republic of Turkey Ministry that COVID-19 has 308069 confirmed cases a with 2.5 % death rate [4].

Because of the high rate of mortality of COVID-19 it has become one of the most important global threats. The COVID-19 has a great deal of importance among many researchers not only from medicine or biology but also from applied sciences as well as governments. Any contribution to theoretical, experimental, or clinical research plays an important role to understand the truth behind that disease, making predictions and, providing insights and strategies to overcome the outbreak which can be used by government authorities.

There are various remarkable studies on mathematical modeling of the first wave of COVID-19 in the time window Dec 2019-July 2020. To the best knowledge of authors, the second wave of COVID-19 has not been declared yet. However, the reported number of diseases increasing after a decrease as a result of the interventions taken can be considered as evidence to accept the second wave. A new fractional COVID-19 mathematical model with a lock-down effect has been studied in [5]. An analysis and estimation of COVID-19 spreading in China, Italy, and France have been studied in [6]. Moreover,

a reduced-space Gaussian process regression model has been proposed in [7]. SEIR model is used to suggest a predictive analysis of the peak outbreak epidemic in South Africa, Turkey, and Brazil by the study of [8]. Authors in [9] propose predictions of various parameters related to COVID-19 and on the number of susceptible, infected, and removed populations in different communities including China, South Korea, the USA, Australia, and India. A mathematical model from a general point of view has been developed based on the Be-CoDiS model given in [10]. However, the obtained model needs several parameters which are not easy to get from available data. A work consisting of the SEIR model and a linear regression model has been conducted by Pandey et al., [11]. The author forecasted the second wave of the spread of the COVID-19 in Iran in [12]. There is a lack but remarkable studies on the modeling of COVID-19 for the case of Turkey. A mathematical model of COVID-19 spread in Turkey and South Africa has been proposed by Atangana and İğret-Araz [13]. Niazkar et al. applied three explicit mathematical prediction models including a recursive-based method, Boltzmann function-based model, and Beesham's prediction model to forecast the COVID-19 outbreak in Iran and Turkey between 15 May to 4 June 2020 in [14]. Furthermore, a SIR model has been studied in [15].

The main aim of this study is to propose a nonlinear regression model for the second wave of COVID-19 for the case of Turkey. To the best knowledge of authors, this is the first time to develop such a model including its all mathematical aspects. Here our purpose is not to underestimate the studies done previously, but to consider the issue from a mathematical point of view with its all details. The current study covers not only a theoretical investigation but also computational simulation.

After presenting a brief introduction, the outline of the paper is organized as follows: Section 2 describes the methodology and analysis of the model proposed in this study. Results and discussions have been provided in Section 3 which is followed by the Conclusion of the study.

2. METHODOLOGY

The idea behind the regression analysis is to describe a function in order to fit the available data set in the "best" possible manner not to match the data exactly. Thus, the main goal of this section is to propose a mathematical model which fits the time series data set provided by the Republic of Turkey Ministry, [4].

Due to the fluctuations in the number of performed tests, the proportions must be studied instead of the number of infected and recovered individuals. In this study, the total number of the population has been assumed as constant in time. Moreover, the stochastic characteristic of the model such as the government measures or the number of contact of infected individuals is excluded for the sake of intelligibility.

Based on the mentioned assumptions and concepts above, let I_P and R_C stand for the ratio of total infected individuals to Turkey's population (assumed approximately 83 million) and the ratio of total recovered individuals to the total confirmed cases, respectively. By reason of emerging the recovered individuals from infected ones, the alteration on this ratio can affect the value of I_P. It is important to emphasize that such a system corresponds to a dynamic system. This means that all variables affecting the systems are also effects each other. However, in this study, we essentially focus on the prediction of infected individuals for the second wave of COVID-19 disease over time. Thus R_C can be considered as an independent variable when I_P is a dependent one. More clearly, I_P is now our dependent variable whereas R_C and t (represents time) independent variables. To understand how these variables are correlated we use Figure 1.
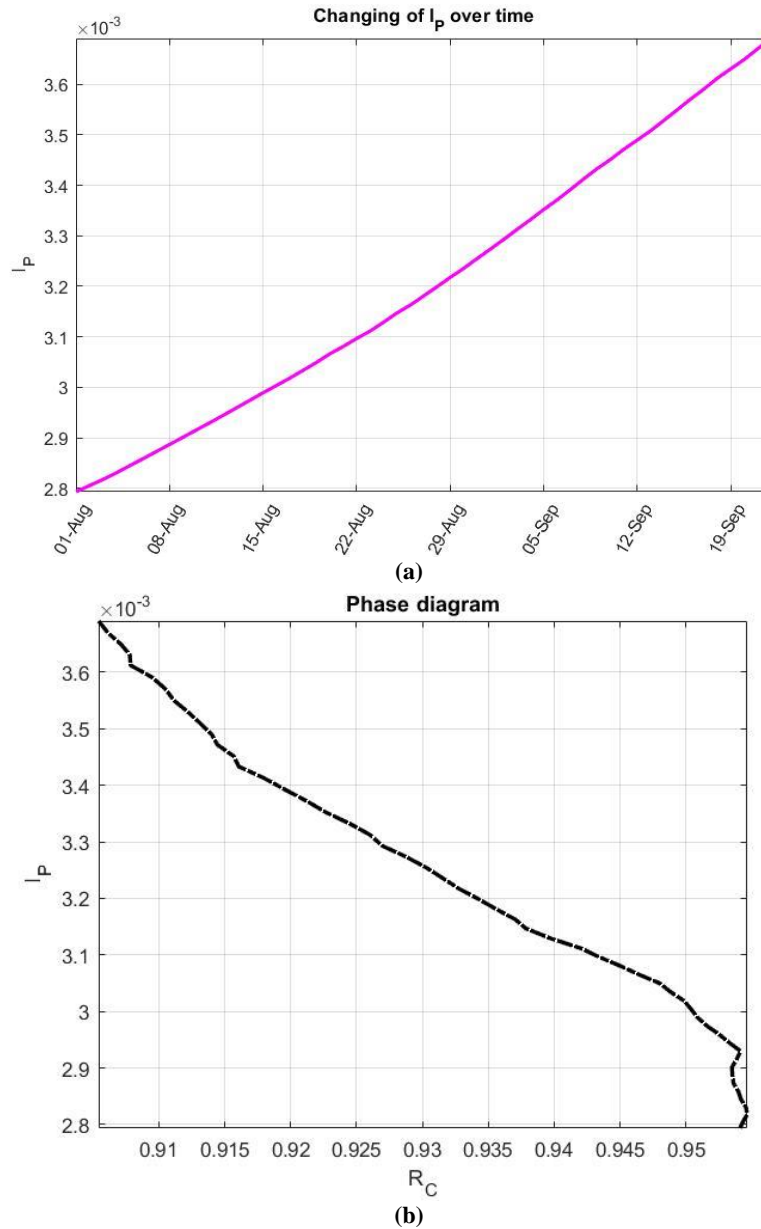
(a)

(b)

Figure 1. The graphical representations of the correlations: (a) the correlations between time versus I_P; (b) the phase diagram of R_C versus I_P.

Figure 1 illustrates the correlations between time versus I_P and the phase diagram of R_C versus I_P. The time period of data is from 1 Aug-22 Sep 2020. Figure 1 exhibits that I_P has almost between a linear and a quadratic shape but taking into account the nonlinear behavior of the real-situation it would be better to consider the curve has the shape of t^β for any $\beta \in \mathbb{R}$. In this context, it is expected that $1 \leq \beta \leq 2$. Moreover, with the help of Figure 1 R_C affects I_P linearly in the negative direction. However, R_C is also a function of t. Thus the usage of one independent variable is better for the simplicity of calculations. The proposed model becomes as follows:

$$I_P(\tau) = \alpha\tau^\beta - \kappa\tau + \delta \tag{1}$$

where τ denotes the rate of change per day. For the sake of simplicity, we assumed that at the initial time, $t = 0$, the approximation, and the data give the same value, that is $\delta = I_P(0)$. This means that the error at the initial time is accepted as zero. To account for this, we define $\tilde{I}_{P,i} := I_{P,i} - I_P(0)$, and Equation (1) becomes

$$\tilde{I}_P(\tau) = \alpha\tau^\beta - \kappa\tau \tag{2}$$

As mentioned earlier, the main idea behind the regression analysis is to get the best curve that interpolates to the raw/exact data. Thus, before training the data we first determine the choices of parameters of α, β, and κ such that Equation (2) fits the data as best as possible. The best fit also means that the minimum error. For that purpose, we define the error function which will be minimized as follows:

$$E(\alpha,\beta,\kappa) = \left(\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{I}_{P,i} - \alpha\tau_i^\beta + \kappa\tau_i\right)^2\right)^{1/2} \tag{3}$$

where $E(\alpha,\beta,\kappa)$, and N stands for the root mean square error (RMSE) and length of the trained data set, respectively. Additionally, $I_{P,i}$ and τ_i denote the value of I_P, and τ at i^{th}day, respectively. After that, the error function given in Equation (3) can be minimized by the virtue of the following systems of three equations:

$$\sum_{i=1}^{N}\left(\tilde{I}_{P,i} - \alpha\tau_i^\beta + \kappa\tau_i\right)\left(-\tau_i^\beta\right) = 0$$
$$\sum_{i=1}^{N}\left(\tilde{I}_{P,i} - \alpha\tau_i^\beta + \kappa\tau_i\right)\left(-\alpha\tau_i^\beta \ln\beta\right) = 0 \tag{4}$$
$$\sum_{i=1}^{N}\left(\tilde{I}_{P,i} - \alpha\tau_i^\beta + \kappa\tau_i\right)(\tau_i) = 0$$

Notice that Equation (4) can be obtained by the standard procedure of the optimization, that is $\nabla E = \left(\frac{\partial E}{\partial\alpha}, \frac{\partial E}{\partial\beta}, \frac{\partial E}{\partial\kappa}\right) = 0$. The authors refer the interested reader to [16] for detailed discussion.

Putting Equation (3) into the algebraic system we have

$$\underbrace{\begin{bmatrix} \sum_{i=1}^{N}\tau_i^{2\beta} & 0 & -\sum_{i=1}^{N}\tau_i^{\beta+1} \\ \sum_{i=1}^{N}\tau_i^{2\beta}\ln\beta & -\kappa\sum_{i=1}^{N}\tau_i^{\beta+1} & 0 \\ \sum_{i=1}^{N}\tau_i^{\beta+1} & 0 & -\sum_{i=1}^{N}\tau_i^2 \end{bmatrix}}_{F(\mathbf{X})} \underbrace{\begin{bmatrix} \alpha \\ \ln\beta \\ \kappa \end{bmatrix}}_{X} - \begin{bmatrix} \sum_{i=1}^{N}\tilde{I}_{P,i}\tau_i^\beta \\ \sum_{i=1}^{N}\alpha\tilde{I}_{P,i}\tau_i^\beta \ln\beta \\ \sum_{i=1}^{N}\tilde{I}_{P,i}\tau_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{5}$$

Due to the nonlinearity of the system Equation (4) is now considered as a root-finding problem. Thus it is solved by a numerical solver using MATLAB software. The following algorithm describes the procedure.

Algorithm 1. The numerical flow to obtain the parameters

Procedure: The parameters to minimize the cumulative error given Equation (2)
Inputs: *insert data set, error*
Define $I = I_{P,i} - I_{P,i}(0)$ *for i=1 to N*
Set tolerance$=10^{-5}$
Initialize for $\mathrm{x} = [\alpha, \beta, \kappa]^T$ *% the initial guess for nonlinear solver*
Define $\mathrm{x}_{old} = [\alpha, \ln\beta, \kappa]^T$
while   error > tolerance do
    $\mathrm{x}_{new} \leftarrow solve\ F(\mathrm{x}_{old})$ *% using a nonlinear solver*
    $error \leftarrow \|\mathrm{x}_{new} - \mathrm{x}_{old}\|_2$
    $\mathrm{x}_{old} \leftarrow \mathrm{x}_{new}$ *% update the solution*
    $\mathrm{x} \leftarrow \left[\mathrm{x}_{new}(1), \exp\left(\mathrm{x}_{new}(2)\right), \mathrm{x}_{new}(3)\right]^T$ *% update the solution*
end while
Outputs: $I_P = I_{P,i}(0) + \mathrm{x}_{new}(1)\tau^{\mathrm{x}new(2)} - \mathrm{x}_{new}(3)\tau$ *% for the infected ratio curve*
        $TI = (83e + 6)I_P$ *% describe the total number of infected individuals*

After the estimated values are computed we evaluate the standard error of estimate, σ_{est}, which tells us straight up how precise the model's predictions are, using

$$\sigma_{est} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

where y_i is raw data, and $\hat{y}_i$ is the predicted data elements, $i = 1, 2, .., N,$ and N is the number of data.

3. RESULTS AND DISCUSSIONS

In Turkey, the first case of COVID-19 was declared on 11th March 2020. The first wave of the disease reached its peak value in one-month duration per day and by the virtue of some interventions provided by the government, it starts to decrease. After a while of the normalization process on 1st June 2020, the number of cases has begun to increase. With the help of the available data set, some fluctuations are observed between June and July. However, a steady increase has been observed as of 1 August 2020.

As mentioned before, the model has been studied to forecast the total number of cases over time. In the current study, we have used the data from 1st August 2020 to 10th September 2020 as our training set and the data from 11th September to 22nd September as the test/evaluation data. The following figure, Figure 2, depicts that the proposed model is in perfect agreement with the available data.
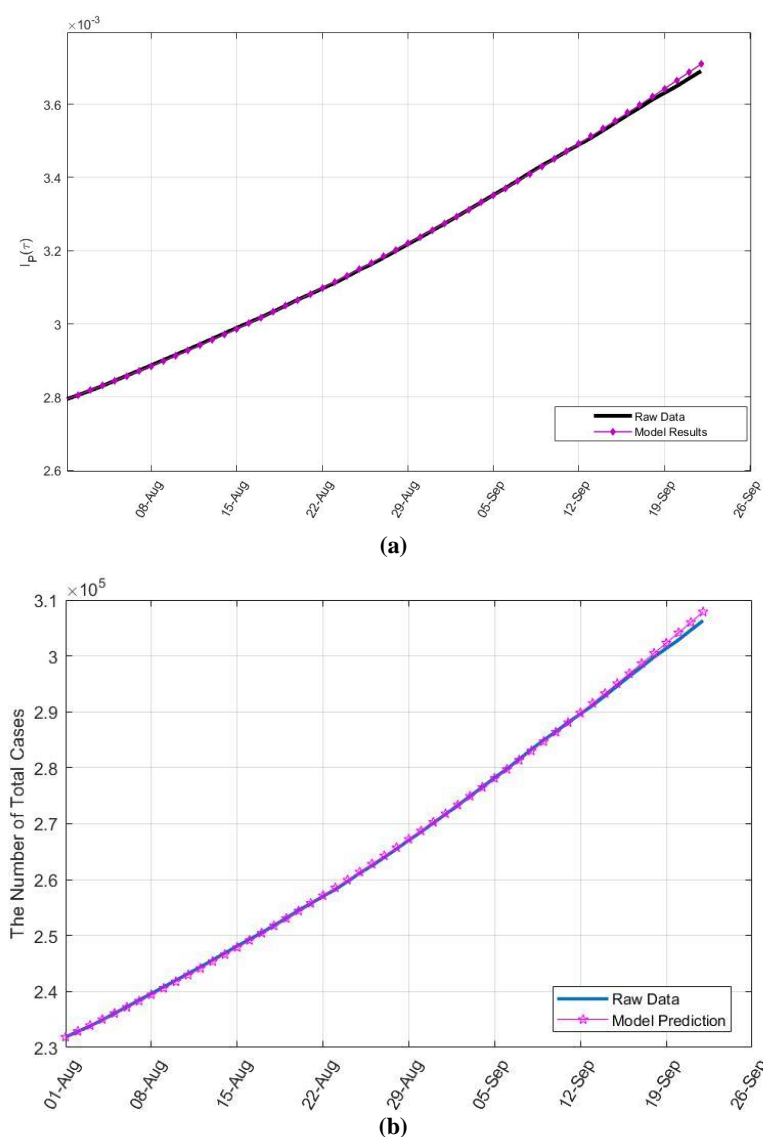

(a)

(b)

Figure 2. The comparison of the proposed model by the raw data: (a)The model results of I_P compared with raw data proportions from 1 Aug -22 Sep 2020; (b) The model estimation for total confirmed cases comparing with raw data from 1Aug-22 Sep 2020.

The obtained results for the proposed model is presented graphically by comparing with the raw data in Figure 2. The parameters for the model are obtained as follows:

$$\alpha = 1.202344e - 07, \beta = 1.960446 \text{ and } \kappa = -1.226419e - 05. \tag{6}$$

Even though the values of α and κ seem without effect, it is crucial to say that these parameters are obtained for the proportions. This means that once the population is taking into account these parameters has an impact on the solution. Furthermore, the value of β is in good accordance with the theoretical expectation. Table 1 lists the comparison of the number of total cases between the proposed model prediction and the raw data set.

Table 1. The Number of Total Confirmed Cases of Raw Data and Model Predictions

Day	Model Prediction	Total Confirmed Cases
11 Sep 2020	288088	288126
12 Sep 2020	289806	289635
13 Sep 2020	291541	291162
14 Sep 2020	293292	292878
15 Sep 2020	295059	294620
16 Sep 2020	296843	296391
17 Sep 2020	298642	298039
18 Sep 2020	300458	299810
19 Sep 2020	302290	301348
20 Sep 2020	304137	302867
21 Sep 2020	306001	304610
22 Sep 2020	307881	306302

In addition to Table 1, Table 2 lists the comparison of the proposed model daily prediction with the available data. It can be seen that the variation in the number of daily confirmed cases is probably because of the number of performed tests. Thus, those values have fluctuations whereas the model prediction does not.

Table 2: The Number of Infected Individuals Model Results and the Raw Data per day

Day	Model Prediction (day-1)	Infected Individuals (day-1)	Performed Tests (day-1)
11 Sep 2020	1702	1671	112.213
12 Sep 2020	1719	1509	98.326
13 Sep 2020	1735	1527	96.097
14 Sep 2020	1751	1716	112.563
15 Sep 2020	1767	1742	110.412
16 Sep 2020	1783	1771	112.645
17 Sep 2020	1799	1648	109.985
18 Sep 2020	1816	1771	111.113
19 Sep 2020	1832	1538	97.416
20 Sep 2020	1848	1519	95.321
21 Sep 2020	1864	1743	112.942
22 Sep 2020	1880	1692	114.311

Furthermore, from 1 August 2020 to 22 September 2020, for Total Confirmed cases, σ_{est} was calculated as 425 individuals with 95 % confidence interval $(\bar{\mu}_P - 1.96 * 425, \bar{\mu}_P + 1.96 * 425)$ where the mean of predicted data was $\bar{\mu}_P = 266203$. Likewise, for Total Death cases, σ_{est} was 55 with 95 % confidence interval $(\bar{\mu}_P - 1.96 * 55, \bar{\mu}_P + 1.96 * 55)$ where $\bar{\mu}_P = 6412$ individuals. Furthermore, σ_{est} was calculated for the Infected Individuals as 103 with 95 % confidence interval $(\bar{\mu}_P - 1.96 * 103, \bar{\mu}_P + 1.96 * 103)$ where $\bar{\mu}_P = 1462$. It is critical to note that the obtained standard errors are quite low when the given number of individuals is compared by the population of Turkey, 83 million. We also note that the smaller the standard error of the estimate is, the more accurate the predictions are.

Because of the theoretical expectation for parameter $1 \le \beta \le 2$, further discussion has been done in Table 3 for comparing the model results with linear and quadratic regression models. For the sake of consistency, after applying a similar analysis the linear and quadratic parameters are obtained. The following table presents the standard error (σ_{est}), and absolute error (AE) which is defined as $\sum_{i=1}^{N=53} |y_i - \hat{y}_i|$ where y_i and $\hat{y}_i$ stand for raw data and the predicted data, respectively.

Table 3: The comparison of errors for various models.

Model	Linear Curve	Quadratic Curve	Proposed Model
σ_{est}	3206	469	425
AE	1.5275e-3	1.7985e-4	1.7086e-4

Table 3 shows the proposed model achieves the best records for the Total Confirmed cases from 1 August 2020 to 22 September 2020. Even though the absolute errors of the quadratic curve and the proposed model seem so close to each other, the recorded minimum error is obtained when $\beta = 1.96$. This a piece of evidence that the obtained numerical results and the theoretical analysis are consistent.

4. CONCLUSION

In this study, a nonlinear regression model has been proposed by taking into account the characteristics of COVID-19. For this purpose, we have utilized the available data set. This novel approach covers the time series from 1 August 2020 to 22 September 2020 on which the increases of the cases have begun. Moreover, the proposed model has been analyzed in detail from a mathematical point of view. The obtained results and simulations have shown a perfect match with the raw data with standard error and with standard deviation. The model has achieved the best records when comparing with the linear and quadratic curves, as well. This is evidence that theoretical analysis plays a significant role in understanding data. According to the model results, the total number of cases at the end of October is expected to become 390100 with 9362 total deaths with a 12 % margin of error unless any measures are taken. Moreover, those numbers will be revised as 473100 cases with 11354 total deaths with around 15 % margin of error and 572700 cases with 13745 deaths with a 18 % margin of errors at the end of November and December, respectively. It is vital to note that these predictions will lose their reliability over time unless the model is renewed.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST

The authors stated that there are no conflicts of interest regarding the publication of this article.

REFERENCES

[1]   Ivorra B, Ferrandez M, Vela-Perez M & Ramos A. Mathematical Modelling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. Commun. Nonlinear Sci. Numer. Simulations, 2020; 88: 105303.

[2]   World Health Organization. Coronavirus disease (COVID-19) outbreak. https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[3]   World Health Organization, 2. (2020). WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/

[4]   Ministry RT. Rebuplic of Turkey Ministry, COVID-19 Information Page. https://covid19.saglik.gov.tr/TR-66122/genel-koronavirus-tablosu.html

[5]   Atangana A. Modelling the spread of COVID-19 with new fractal-fractional operators. Can the lockdown save mankind before vaccination? Chaos, Solitons, Fractals, 2020; 136: 109860.

[6]   Fanelli D & Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy, and France. Chaos Solitons Fractals, 2020; 134: 109761.

[7]   Velásquez R & Vanessa J. Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. Chaos Solitons Fractals, 2020; 136: 109924.

[8]   Djilalia S & Ghanbari B. Coronavirus pandemic: A predictive analysis of the peak outbreak epidemic in South Africa, Turkey, and Brazil. Chaos Solitons Fractals., 2020; 138: 109971.

[9]   Cooper I, Mondal A & Antonopoulos CG. A SIR model assumption for the spread of COVID-19 in different communities. Chaos, Solitons and Fractals, 2020; 139: 110057.

[10]   Ivorra B Ngom D & Ramos AM. Be-CoDiS: A Mathematical Model to Predict the Risk of Human Diseases Spread Between Countries-Validation and Application to the 2014-2015 Ebola Virus Disease Epidemic. Bull Math Biol, 2015; 77 (9): 1668-1704.

[11]   Pandey G, Chaudhary P, Gupta R & Pal S. SEIR and Regression Model based COVID-19 outbreak predictions in India. Medrxiv, 2020. DOI: https://doi.org/10.1101/2020.04.01.20049825

[12]   Ghanbari B. On forecasting the spread of the COVID-19 in Iran: The second wave. Chaos, Solitons and Fractals, 2020; 140: 110176.

[13]   Atangana A. & Igret-Araz S. Mathematical model of COVID-19 spread in Turkey and South Africa: Theory, methods and applications: MedRxiv 2020. DOI: https://doi.org/10.1101/2020.05.08.20095588

[14]   Niazkar M, Eryılmaz Türkkan G, Niazkar HR, Türkkan YA. Assessment of Three Mathematical Prediction Models for Forecasting the COVID-19 Outbreak in Iran and Turkey. Computational and Mathematical Methods in Medicine, 2020; vol. 2020: 7056285.

[15] Ozdinc M, Senel K, Ozturkcan S & Akgul A. Predicting the Progress of COVID-19: The Case for Turkey. Turkiye Klinikleri Journal of Medical Sciences, 2020; 40 (2): 117-119.

[16] Burden RL, Faires JD. Numerical Analysis, 9th Ed. Canada, USA: Cengage Learning, 2010.