Contents lists available at ScienceDirect

Journal of Infection and Public Health

journal homepage: http://www.elsevier.com/locate/jiph

COVID-19 prevalence forecasting using Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN): Case of Turkey

Gülhan Toğa, Berrin Atalay, M. Duran Toksari*

Erciyes University, Engineering Faculty, Industrial Engineering Department, Kayseri, Turkey

ARTICLE INFO

Article history:
Received 1 September 2020
Received in revised form 23 April 2021
Accepted 27 April 2021

Keywords:
ARIMA
ANN
COVID-19
Turkey
Forecasting

ABSTRACT

A local outbreak of unknown pneumonia was detected in Wuhan (Hubei, China) in December 2019. It is determined to be caused by a severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) and called COVID-19 by scientists. The outbreak has since spread all over the world with a total of 120,815,512 cases and 2,673,308 deaths as of 16 March 2021. The health systems in the world collapsed in many countries due to the pandemic and many countries were negatively affected in the social life. In such situations, it is very important to predict the load that will occur in the health system of a country. In this study, the COVID-19 prevalence of Turkey is inspected. The infected cases, the number of deaths, and the recovered cases are predicted with Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN) in Turkey. The techniques are compared in terms of correlation coefficient and mean square error (MSE). The results showed that the used techniques used are very successful in the estimation of prevalence in Turkey.

Introduction

The COVID-19 pandemic is the virus outbreak that occurred in Wuhan, the capital of China's Hubei region on December 1, 2019. The newly observed coronavirus called SARS-CoV-2, similar to pneumonia was diagnosed. Infectiousness of the COVID-19, transmission of the virus from person to person, grew in mid-January 2020. In a very short time, virus cases in various countries in Europe, North America, and Asia-Pacific started to be reported. A global pandemic was declared by the World Health Organization (WHO) on March 11, 2020. On March 13, 2020, WHO reported that the new epicenter of the coronavirus pandemic is Europe [1].

After the COVID-19 outbreak spreading across all over the world, the first detected COVID-19 cases were announced on 10 March 2020, by the Ministry of Health in Turkey. The first virus-related death in Turkey occurred on March 15, 2020. The total number of patients infected with the coronavirus was announced as 2,911,642 while 29,623 people died as of March 16, 2021, in Turkey. 2,734,862 patients recovered and gain immunity as of 16 March, 2021.

While the COVID-19 pandemic spreads over the world, literature has grown rapidly with the studies of scientist. Therefore, COVID-19 studies have received much attention. There are many studies in the literature about the medical aspect of COVID-19 [2–4]. On the other hand, sociological [5], economical [6], and statistical [7] inspections of COVID-19 are made by the researchers in many studies. Statistical studies generally focus on the country-based forecasting of the pandemic. Besides the many countries [8–11], Turkey is also investigated in some studies [12–14]. Statistical and forecasting studies used different techniques such as time-series analyses, data mining techniques, growth models, nonlinear regression analysis, epidemiological models, and artificial intelligence (AI) techniques. One of the most effective time-series-based methods is ARIMA in the COVID-19 forecasting studies. Many country-based applications used ARIMA in the COVID-19 literature [8,9,11,15–19]. Furthermore, ANN has also been used for predicting the prevalence of COVID-19 in many studies and reported as a successful tool for prevalence prediction [10,20–23].

In this paper, we inspected the dynamics of COVID-19 prevalence in Turkey. Infected cases, number of deaths, and recovered cases are handled, and prediction models are built by using ARIMA and ANN. This paper is organized as follows: The first section gives a brief review of COVID-19 and literature on forecasting the prevalence of the COVID-19 pandemic. The second section examines the

* Corresponding author.
E-mail address: dtoksari@erciyes.edu.tr (M.D. Toksari).

https://doi.org/10.1016/j.jiph.2021.04.015

materials and methods used for the prediction of the prevalence of pandemic. Results and discussions are given in the third section. Our conclusions are drawn in the final section.

Materials and methods

This section presents two approaches such as ARIMA and ANN for COVID-19 prevalence forecasting of Turkey in this study.

ARIMA method

ARIMA model, also known as Box and Jenkins, is one of the statistical methods used for future prediction. The Box–Jenkins method is used in the future prediction of univariate time series. It shows a systematic approach in establishing future prediction models of discrete and stationary time series consisting of observation values obtained with equal time intervals. The series consisting of observation values obtained with equal time intervals are important assumptions of the Box–Jenkins method which is discrete and stationary [24]. It is also a very effective tool in estimating time-series data. ARIMA method combines AR (autoregressive) and MA (moving averages) to analyze data. ARIMA models are used for stationary time series. Stabilization of the data is carried out by taking the difference in the I (Integrated-d) process. If the degree of autoregression parameter is p, the degree of difference parameter is d, and the degree of moving average parameter is q this model is called Autoregressive Integrated Moving Average model in degrees $(p,\ d,\ q)$ and it is written as ARIMA $(p,\ d,\ q)$ [24].

The general expression of an ARIMA $(p,\ d,\ q)$ model is as follows:

$$w_t = Ø_1 w_{t-1} + Ø_2 w_{t-2} + \ldots + Ø_p w_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \quad (1)$$

If the primary differences $(d = 1)$ make the stationary series, the difference operator will be as follows:

$$\nabla x_t = w_t = x_t - x_{t-1} \quad (2)$$

∇ = difference operator,
d = degree of difference,
$\{w_t\}$ = differenced series.

The number of parameters to be calculated in the general ARIMA $(p,\ d,\ q)$ model used in the future estimation of series that do not show seasonal fluctuations is as much as in ARMA $(p,\ q)$. In ARIMA $(p,\ d,\ q)$ model, p or q can be zero. In this case, the model is reduced to either the AR $(d,\ p)$ or MA $(d,\ q)$ model types.

Artificial Neural Networks

ANN is one of the highly effective and successful data mining techniques in the literature. ANN is an information processing method inspired by the human brain. A brain learns from human experiences and ANN mimics the brain while processing the data. It is classified as supervised or unsupervised learning according to the knowledge of the output variable values. Generally, ANN consists of some basic elements: input, hidden, and output layers. An input layer is the information provider of the networks. The hidden layer constructs the nonlinear relations between input(s) and output(s) by adjusting weights and this step is called learning. Layers consist of different numbers of neurons and these neurons process the data via activation functions. On the other hand, the output layer gives the forecasting information.

It is proper to use ANN if there is no theoretical information about the functional form of the model or the nonlinear structure of the model. This leads us it is not a model-based technique; it is a data-based technique. The general architecture of ANN is given in Fig. 1. Data are generally split into categories for training, testing, and validation purposes. In the training step, a network learns from the data. Stopping the training process is achieved by the validation step, while the prediction ability of a trained ANN is judged in the testing step [25].

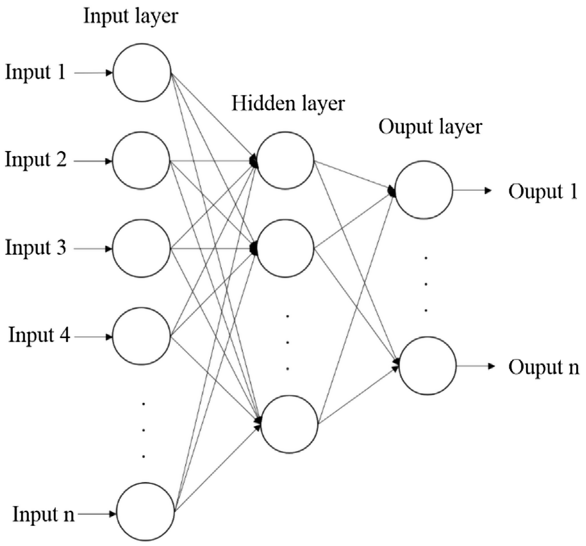

Fig. 1. ANN architecture.

A special type of feedforward neural network called multilayer perceptron (MLP) is used in this study. The MLP is the most widely used ANN model and generally contains one input layer, one output layer, and one or more hidden layers (Basheer and Hajmeer, 2000). Different training algorithms are used for MLP neural networks. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) is one of the training algorithms usually used for nonlinear least squares is presented and the modified backpropagation algorithm is combined with the BFGS algorithm [26]. Therefore, BFGS is preferred in this study.

Results and discussion

The daily confirmed COVID-19 data from March 31, 2020 to March 16, 2021 are retrieved from the website of Turkish Ministry of Health (daily data not announced by the government has been neglected). On the other hand, population related data are gathered from the website of Turkish Statistical Institute. Only the patients that were confirmed by laboratory tests as positive are considered as infected cases by the Turkish government and we use these data for analyzes. In this study, no primary data collection is undertaken, no patient or public was involved in the study. By the way, we do not need any formal ethical assessment or informed consent. All anonymized data are collected from the official websites.

Results of the ARIMA

In our study, the ARIMA method is used for the prediction of the daily number of infected cases, the daily number of deaths, and the number of recovered cases. ARIMA model cannot be generated models for multiple outcomes. Therefore, we structure three different ARIMA models by using Minitab 17.3.1 software. As the first step of the ARIMA, the stationary condition of the time series is checked. And it is seen in Fig. 2 that our data are not stationary.

The daily number of infected cases is checked on the ACF graph in Fig. 3. In the ACF graph, it is seen that the daily number of the infected case has a serial trend. Therefore, data are preprocessed by

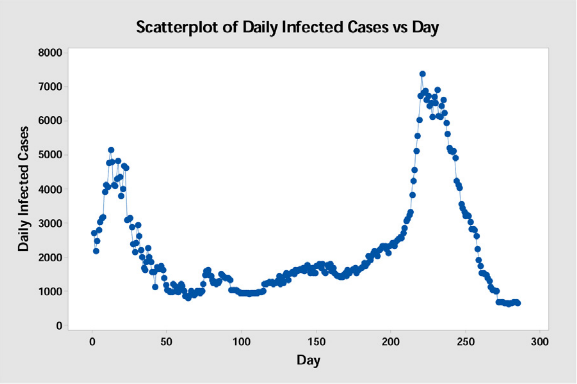

Fig. 2. Daily number of infected cases vs. day graph.

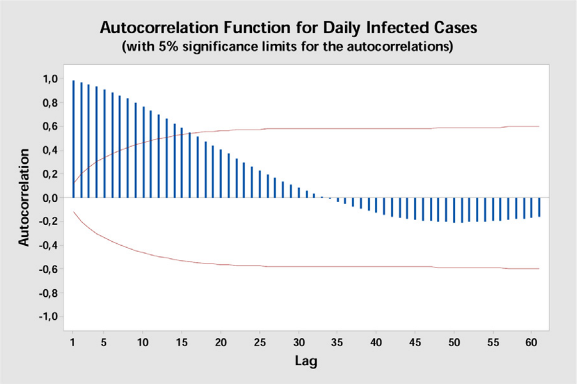

Fig. 3. Daily number of infected cases Autocorrelation Coefficient Function (ACF) graph.

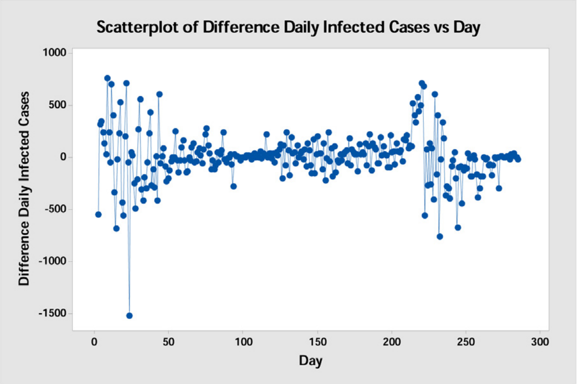

Fig. 4. Daily number of infected cases vs. day graph with differenced data.

differencing. The stationary condition has been provided by differenced time series as seen in Fig. 4.

Deciding on the values of p, q and d is the crucial point in the ARIMA model and these parameters affect the performance of the ARIMA model. In this study, ARIMA models are created with different combinations of p, q, and d values, and their performances are compared. The ACF and Partial Autocorrelation Coefficient Function (PACF) graphs are plotted to choose the best performing ARIMA model in Figs. 5 and 6. According to the ACF and PACF graph, high autocorrelation is observed among the data.

ARIMA $(1, 1, 0)$ model gives the best forecasting results for the daily number of infected cases. The p value obtained as 0.000 for the daily number of infected cases in the level of 5% significance. The same procedures are applied to forecasting the number of daily

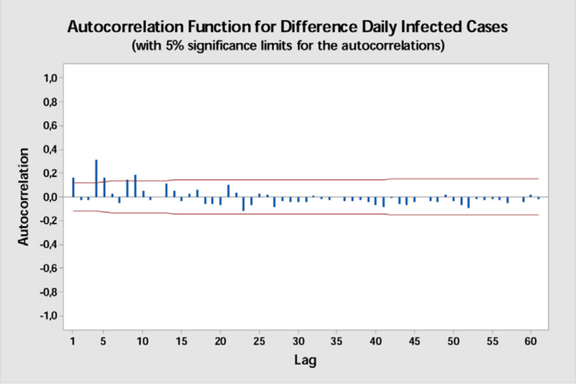

Fig. 5. Differenced daily number of infected cases ACF graph.

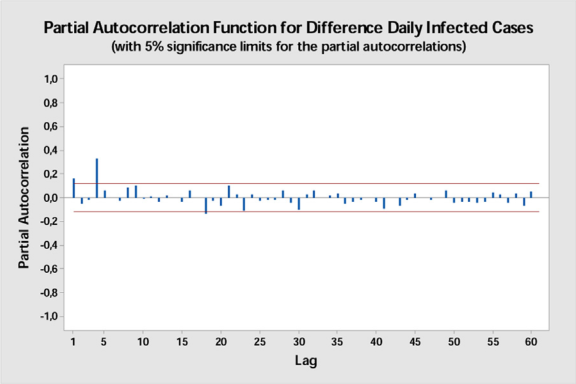

Fig. 6. Differenced daily number of infected cases PACF graph.

deaths and the number of daily recovered cases. Table 1 shows the Pearson correlation (R) value, sum of square error (SSE), MSE, and p values obtained for all estimation parameters.

As can be seen from Table 1, the highest correlation value is obtained from the daily number of recovered cases. According to SSE and MSE values, it is seen that the minimum errors are observed in the estimation of the daily number of deaths.

Results of ANN

We have a sample size of 285 days. The sample size is directly related to the generalization ability of ANN. ANN, generally, converges at the local minima with small sample sizes and yields poor generalization [27]. To overcome this issue, the data set is divided into three samples as 70% training, 15% testing, and 15% validation in this study. With the validation process, the generalization ability of constructed networks is tested. Each network is trained during 200 cycles and stopped when a 0.0000001 change in the error is occupied and 500 different network architecture is run. ANN structures are evolved from regression-based time series analysis models in our study.

Inputs and outputs of the model are defined as below:

- Susceptible cases
- Days
- Curfews
- Laboratory tests

Outputs:

- Daily number of infected cases
- Daily number of recovered cases

Table 1
Parameters and results.

	ARIMA (p,d,q)	R	SSE	MSE	p-Value
Daily number of infected cases	ARIMA (1,1,0)	0.987	15,606,683	55,343	0.000
Daily number of deaths	ARIMA (0,1,1)	0.996	13,600.8	48.2	0.000
Daily number of recovered cases	ARIMA (1,1,1)	0.998	815,523,606	2,912,584	0.000

Table 2
Best network architecture.

	Training perf.	Testing perf.	Validation perf.	Training algorithm	Error function	Hidden activation	Output activation
MLP 5-10-3	0.98	0.98	0.98	BFGS	SSE	Hyperbolic tangent	Logistic

Table 3
SSE and MSE of ANN.

	Training error	Testing error	Validation error
SSE	1,061,922	1,160,705	1,628,951
MSE	3726.04	4072.65	5715.62

Table 4
Pearson correlation values of parameters.

R values of MLP 5-10-3	Train	Test	Validation
Daily number of infected cases	0.98	0.97	0.98
Daily number of deaths	0.99	0.99	0.99
Daily number of recovered cases	0.97	0.97	0.98

- Daily number of deaths

To determine whether the input parameters are statistically significant for the ANN model, we check the p-values of input variables. p-Values of input parameters are found as 0.048, 0.000, 0.000 and 0.000 for susceptible cases, days, curfews and laboratory tests, respectively. Our finding support that all input parameters are statistically significant for the ANN model with p-values smaller than 0.05.

We calculate the susceptible case number for each day after the first recovered case was reported in Turkey. It is updated by the formulation *Susceptible = Population − Mortality − Recovered cases*. Non-pharmaceutical interventions such as school closures, travel restrictions, curfews, and quarantines are applied in Turkey during COVID-19. Travel restrictions, school closures, and quarantine policies are regularly applied after the first case was reported in Turkey. However, curfews varied during the pandemic. This is an important parameter that caused fluctuations in the number of infected people. Curfews are coded as 1 for the days applied and 0 for the other days. One of the most important issues of forecasting accuracy is that cases are confirmed and reported after the laboratory tests give positive results. By the way, this is another critical parameter for our model.

On the other hand, multicollinearity among independent variables is an important assumption of regression-based approaches. To check this assumption, we analyze independent data to detect multicollinearity. The most common approach to detect multicollinearity is that of the variance inflation factor (VIF). Depending on the rules of 4 or 10, multicollinearity among independent variables can be a possible or serious problem [28]. VIF values for our independent variables range between 1–2; therefore we will not discuss the multicollinearity problem among independent variables. ANN analyzes are carried out using the data mining module of the STATISTICA 10.0 software package.

The best network architecture is given in Table 2 and the SSE of the training, testing, and validation steps are given in Table 3.

Pearson Correlation coefficients are given in Table 4. As seen in Table 4, the daily number of deaths, the daily number of recovered cases, and the daily number of infected cases have high correlation coefficient values and this indicates that the model developed has an acceptable generalization capability and accuracy to predict the prevalence of COVID-19 pandemic in Turkey.

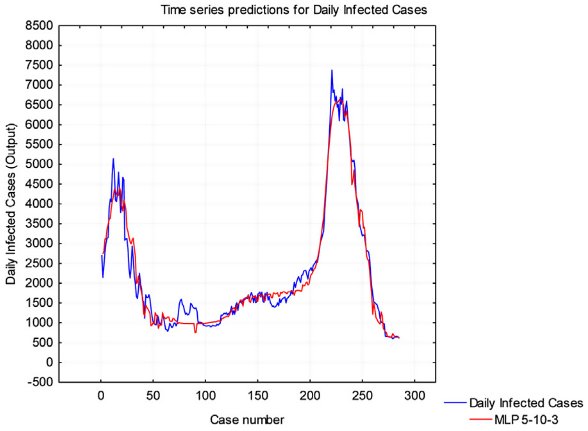

Fig. 7. Time series prediction for infected cases.

As depicted in Table 1, the best network is a multilayer perceptron network consists of five input neurons (curfews are considered as 2 different neurons because of the categorical structure of curfew data), ten neurons with a hidden layer, and three output neurons. The training algorithm is selected as BFGS. On the other hand, activation functions are selected as hyperbolic tangent and logistic functions for the hidden layer and output layer, respectively. Furthermore, the selected network accurately predicts the daily number of infected cases, daily number of deaths, and daily number of recovered cases as seen in Tables 3 and 4.

High correlation coefficients may suspect about linear relation between data or poor generalization ability of the developed network. However, as can be seen from Figs. 7 to 9 , the developed model is very successful in nonlinear estimation because the predicted and actual values of output curves overlap in the graphs. Figs. 7–9 give the time series predictions for 3 outputs.

Conclusion

The effect of the COVID-19 outbreak is growing steadily in the whole world. It becomes very important to forecast the prevalence of the pandemic for the health systems of countries. Accurate forecasting will be an insight into strengthening health systems and resource reallocation. In this manner, reliable prediction of the COVID-19 pandemic enables rapid responses, event-based political decisions, and to predict the future of the pandemic. Thereby, minimization of deaths and health system-caused failures is provided.

Time-series-based models such as ARIMA and ANN are very efficient tools for outbreak analysis. This study predicts the daily

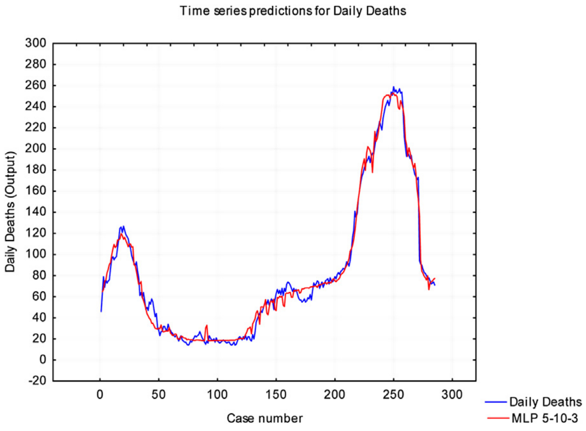

Fig. 8. Time series prediction for the daily number of deaths.

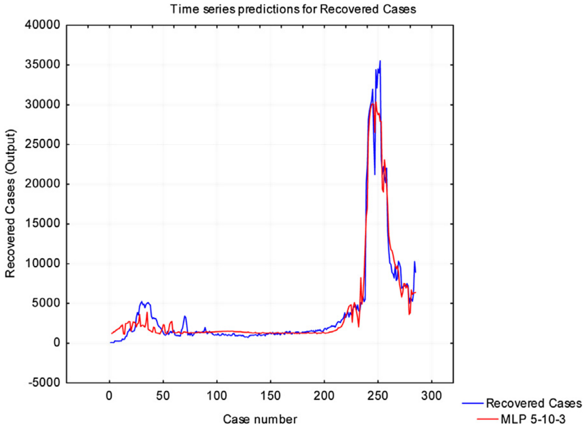

Fig. 9. Time series prediction for recovered cases.

number of infected cases, daily number of deaths, and daily number of recovered cases for Turkey between 31 March–16 March by ANN and ARIMA models. We compared two techniques with some statistical indicators such as MSE and SSE. Following the results, ARIMA and ANN have almost the same forecasting performance. Consequently, ARIMA has high prediction accuracy in this study. R values are very high for predicting prevalence. Three different ARIMA models are developed by using different p, d, and q values. On the other hand, ARIMA has no ability to estimate multiple outputs simultaneously while ANN can construct models that can estimate three variables at the same time at an acceptable prediction level. Additionally, this study has highlighted the success of using artificial intelligence techniques in the estimation of pandemics.

For more precise estimation, data should be updated in real-time and new parameters that will affect the prevalence of pandemic should be taken into account. Vaccination studies have been started on 14 January 2021 in Turkey and it has been considered that the prevalence of pandemics will be affected by vaccination. Therefore, including the vaccination data explained by authorities in the study will be very effective for predicting the prevalence of the pandemic in Turkey in future works.

Funding

No funding sources.

Competing interests

None declared.

Ethical approval

Not required.

References

[1] Vashist SK. In vitro diagnostic assays for COVID-19: recent advances and emerging trends. Diagnostics 2020;10(4):202, http://dx.doi.org/10.3390/diagnostics10040202.

[2] Bai Y, Yao L, Wei T, Tian F, Jin D-Y, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA 2020;323(14):1406–7, http://dx.doi.org/10.1001/jama.2020.2565.

[3] Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ. COVID-19: consider cytokine storm syndromes and immunosuppression. Lancet (London, England) 2020;395(10229):1033–4, http://dx.doi.org/10.1016/S0140-6736(20)30628-0.

[4] Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun 2020;109:102433, http://dx.doi.org/10.1016/j.jaut.2020.102433.

[5] Lancker WV, Parolin Z. COVID-19, school closures, and child poverty: a social crisis in the making. Lancet Public Health 2020;5(5):e243–4, http://dx.doi.org/10.1016/S2468-2667(20)30084-0.

[6] Fernandes N. Economic effects of coronavirus outbreak (COVID-19) on the world economy (SSRN Scholarly Paper ID 3557504). Soc Sci Res Netw 2020, http://dx.doi.org/10.2139/ssrn.3557504.

[7] Roser M, Ritchie H, Ortiz-Ospina E, 45 Coronavirus disease (COVID-19) — statistics and research; 2020.

[8] Al-qaness MAA, Ewees AA, Fan H, Abd El Aziz M. Optimization method for forecasting confirmed cases of COVID-19 in China. J Clin Med 2020;9(3):674, http://dx.doi.org/10.3390/jcm9030674.

[9] Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 2020;729:138817, http://dx.doi.org/10.1016/j.scitotenv.2020.138817.

[10] Moftakhar L, Seif M, Safe MS. Exponentially increasing trend of infected patients with COVID-19 in Iran: a comparison of neural network and ARIMA forecasting models. Iran J Public Health 2020;49(Supple 1):92–100.

[11] Perone G, ArXiv:2004.00382 [q-Bio, Stat] An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy; 2020 http://arxiv.org/abs/2004.00382.

[12] Arslan S, Ozdemir MY, Ucar A. Nowcasting and forecasting the spread of COVID-19 and healthcare demand in Turkey, a modelling study [Preprint]. Public Global Health 2020, http://dx.doi.org/10.1101/2020.04.13.20063305.

[13] Aslan Ibrahim H, Demir M, Wise MM, Lenhart S, 2020.04.11.20061952 Modeling COVID-19: forecasting and analyzing the dynamics of the outbreak in Hubei and Turkey. MedRxiv; 2020 https://doi.org/10.1101/2020.04.11.20061952.

[14] Özdinç M, Şenel K, Öztürkcan S, Akgül A. Predicting the progress of COVID-19: the case for Turkey. Turk Klin J Med Sci 2020;40(2):117–9, http://dx.doi.org/10.5336/medsci.2020-75741.

[15] Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief 2020;29:105340, http://dx.doi.org/10.1016/j.dib.2020.105340.

[16] Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos Solitons Fractals 2020;135:109850, http://dx.doi.org/10.1016/j.chaos.2020.109850.

[17] Dehesh T, Mardani-Fard HA, Dehesh P, 2020.03.13.20035345 Forecasting of COVID-19 confirmed cases in different countries with ARIMA models. MedRxiv; 2020 https://doi.org/10.1101/2020.03.13.20035345.

[18] Ding G, Li X, Shen Y, Fan J. Brief analysis of the ARIMA model on the COVID-19 in Italy. MedRxiv; 2020, http://dx.doi.org/10.1101/2020.04.08.20058636, 2020.04.08.20058636.

[19] Gupta R, Pal SK. Trend analysis and forecasting of COVID-19 outbreak in India. MedRxiv; 2020, http://dx.doi.org/10.1101/2020.03.26.20044511, 2020.03.26.20044511.

[20] Distante C, Pereira IG, Goncalves LMG, Piscitelli P, Miani A. Forecasting Covid-19 outbreak progression in Italian regions: a model based on neural network training from Chinese data. MedRxiv; 2020, http://dx.doi.org/10.1101/2020.04.09.20059055, 2020.04.09.20059055.

[21] Ghazaly NM, Abdel-Fattah MA, El-Aziz AAA. Novel coronavirus forecasting model using nonlinear autoregressive artificial neural network. Int J Adv Sci Technol 2020;29(5):19.

[22] Hasan N. A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model. Internet Things 2020;11:100228, http://dx.doi.org/10.1016/j.iot.2020.100228.

[23] Tamang SK, Singh PD, Datta B. Forecasting of Covid-19 cases based on prediction using artificial neural network curve fitting technique. Global J Environ Sci Manag 2020;6(Special Issue (Covid-19)), http://dx.doi.org/10.22034/GJESM.2019.06.SI.06.

[24] Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. John Wiley & Sons; 2015.

[25] Yaghini M, Khoshraftar MM, Fallahi M. A hybrid algorithm for artificial neural network training. Eng Appl Artif Intell 2013;26(1):293–301, http://dx.doi.org/10.1016/j.engappai.2012.01.023.

[26] Nawi NM, Ransing MR, Ransing RS. An improved learning algorithm based on the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method for back propagation neural networks. Sixth International Conference on Intelligent Systems Design and Applications, vol. 1; 2006. p. 152–7, http://dx.doi.org/10.1109/ISDA.2006.95.

[27] Mao R, Zhu H, Zhang L, Chen A. A new method to assist small data set neural network learning. Sixth International Conference on Intelligent Systems Design and Applications, vol. 1; 2006. p. 17–22, http://dx.doi.org/10.1109/ISDA.2006.67.

[28] O'brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant 2007;41(5):673–90.