Iran J Public Health, Vol. 49, Suppl.1, 2020, pp.92-100

Original Article

Exponentially Increasing Trend of Infected Patients with COVID-19 in Iran: A Comparison of Neural Network and ARIMA Forecasting Models

Leila MOFTAKHAR [1], *Mozhgan SEIF [2], *Marziyeh Sadat SAFE [3]

1. *Student Research Committee, Department of Epidemiology, Shiraz University of Medical Sciences, Shiraz, Iran*
2. *Department of Epidemiology, Faculty of Biostatistics, School of Health, Shiraz University of Medical Sciences, Shiraz, Iran*
3. *Seyed-al-Shohada Hospital, Jahrom University of Medical Sciences, Jahrom, Iran*

***Corresponding Author:** Email: marziyehsafe@gmail.com

(Received 05 Mar 2020; accepted 14 Mar 2020)

Abstract

Background: The outbreak of COVID-19 is rapidly spreading around the world and became a pandemic disease. For help to better planning of interventions, this study was conducted to forecast the number of daily new infected cases with COVID-19 for next thirty days in Iran.

Methods: The information of observed Iranian new cases from 19th Feb to 30th Mar 2020 was used to predict the number of patients until 29th Apr. Artificial Neural Networks (ANN) and Auto-Regressive Integrated Moving Average (ARIMA) models were applied for prediction. The data was prepared from daily reports of Iran Ministry of Health and open datasets provided by the JOHN Hopkins. To compare models, dataset was separated into train and test sets. Mean Squared Error (MSE) and Mean Absolute Error (MAE) was the comparison criteria.

Results: Both algorithms forecasted an exponential increase in number of newly infected patients. If the spreading pattern continues the same as before, the number of daily new cases would be 7872 and 9558 by 29th Apr, respectively by ANN and ARIMA. While Model comparison confirmed that ARIMA prediction was more accurate than ANN.

Conclusion: COVID-19 is contagious disease, and has infected many people in Iran. Our results are an alarm for health policy planners and decision-makers, to make timely decisions, control the disease and provide the equipment needed.

Keywords: COVID-19; Forecast; Artificial neural network; Iran

Introduction

Several cases of severe acute respiratory syndrome with unknown etiology were reported in Dec 2019 in Wuhan City, China (1-5). The coronavirus, a large family of viruses, was the main cause of this outbreak (6-8). Two popular types of coronaviruses are called SARS-CoV-1 and MERS-CoV. They have created two outbreaks in 2003 and 2012, respectively (9). The novel coronavirus is the third type of this family that created a huge pandemic and introduced as COVID-19 by WHO. The origin of this virus is not yet known, but it's more likely to be a bat (10, 11). This disease has an incubation period of over 14 d, its mortality rate is between 2%-3% (12) and is transmitted through respiratory droplets and contact with contaminated surface (9).

COVID-19 spread very fast in China (3, 13), and the world (4). Therefore, the total number of confirmed cases and deaths from this virus 800023 and 38748 people in the world by Mar 31, 2020, and infected more than 190 countries (14). Iran was seen as the first case of COVID-19 in Qom on Feb 19, 2020. Then the disease spread very fast throughout the country (15). The total number of confirmed patients and death in Iran was 44605 and 2898, respectively on Mar 31,2020 (16).

Although, the government has implemented preventive strategies at the beginning of the outbreak, the number of new cases increased and has created a serious concern for people and health policy makers. Because the disease is highly contagious, the government and the health system must be prepared to prevent and counteract it. Therefore, being aware of trend of disease helps to make decisions about preventive interventions. Modeling to predict the number of new cases in the next days is one way that reveals the trend of disease. Artificial Neural Networks (ANN) and Auto-Regressive Integrated Moving Average (ARIMA) are the two most popular class of models for trend modeling and predicting time-series data (17) Although ARIMA has been in use for forecasting infectious disease since past years (18), ANN has been recently known as powerful nonlinear regression techniques(19); and due to its ability for time series forecasting, it has been widely applied (20). ANN is a member of machine learning algorithms. There are many studies confirmed the superior performance of machine learnig algorithms in comparison to more customary models (21). However, none of ARIMA and ANN has been definitively proven to be more precise than the other for different medical fields; and therefore studies continue to compare them (17). This comparison is also continued in this study to determine the most accurate model for forecasting the spreading trend of Coronavirus.

However with the most precise forecasting model since the spread of coronavirus depends on several factors, including environmental factors and personal behavior, quarantine, therefore modeling can not predict the precise number of cases, but they help to make better decisions by health policy makers. Therefore, this study was conducted to forecast daily new cases that infected with COVID-19 for next days in Iran, with identifying the most accurate model for forecasting it.

Methods

This time-series study was conducted to predict the number of new cases infected with COVID-19 in Iran, until 29 Apr 2020. The daily-confirmed cases of COVID-2019 from Feb 19 to Mar 30, 2020, in Iran, were extracted from the daily reports of Ministry of Health and Medical Education of Iran, and open datasets provided by the JOHN Hopkins University.

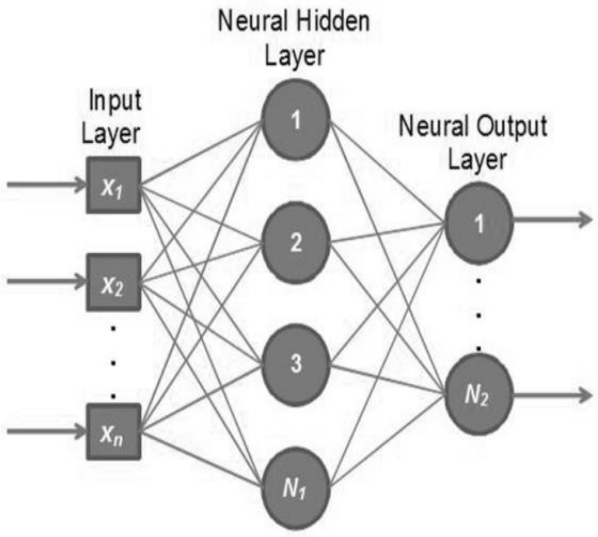
Model introducing

Auto-Regressive Integrated Moving Average (ARIMA)

ARIMA (p, d, q) is simultaneous fit of other two models including Auto-Regressive (p), Moving Average (q) and seasonal autoregressive integrated moving average (SARIMA) model (22, 23). If necessary, ARIMA use differencing (d) to change time series data into stationary ones (17). Box-cox is another popular transformation recommended to improve ARIMA fitting (24). The model goodness of fit is usually assessed through inspection of the residuals; in the way that in a good fit, the plot of residuals versus the order of observations should not depict any trend and the residuals should randomly be scattered around the zero line. The normality of residuals would intuitionally be checked through Normal Probability Plot (NPP) in addition to histogram of residuals; however, normality could analytically be inspected by Shapiro-Wilk test. Autocorrelation and Partial Autocorrelation Functions (ACF & PACF) plots of residuals are other useful tools for goodness of fit assessment. Finally, Residuals could also be tested to be stationary by Box-Ljung.

Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is an extension of Generalized Linear Models (GLM) (25). This data mining algorithm is so popular for modeling nonlinear associations (26). ANN is comprised of three layers: an input, output and hidden layer(s). Each layer is formed from Neurons and Synapses (Fig. 1). The neurons in the input layer are previous observations used for forecasting future values in the output layer. Other layers within input and output are called hidden layers (27-29). Following is an ANN plot with N_1 and N_2 neurons respectively in input and output layers.



Model Implementation

ARIMA and ANN were used as the two most popular class of models for forecasting time series data to forecast the number of newly infected patients. ANN was trained with three hidden layer each of them containing 10 neurons. The number of repetitions for this algorithm was set to be 300. Wherever it was needed, Box-Cox transformation was used to provide normal observations. The goodness of fit for models was assessed through the following plots regarding the residuals: residuals versus observation order, NPP, histogram, ACF and PACF. Residuals were also tested to be stationary by Box-Ljung. For model comparison, the dataset was split into train and test sets respectively including 19th Feb to 24th Mar (35 d) and 25th Mar to 30th Mar (6 d). ARIMA model and ANN algorithm were compared according to Mean Squared Error (MSE) and Mean Absolute Error (MAE) criteria (18), as follows:

$$MSE = \frac{1}{6}\sum_{t=1}^{6}(y_t - \hat{y}_t)^2$$

$$MAE(\%) = \frac{1}{6}\sum_{t=1}^{6}\frac{|y_t - \hat{y}_t|}{y_t} \times 100$$

Finally, data analysis was conducted using 'forecast' and 'nnfor' packages from R software. The significance level was set 0.05.

Results

The observed number of new infected cases from 19th Feb to 30th Mar 2020 is shown in Fig. 2. This Figure also displays the predicted number and its 95% confidence intervals for the next thirty days, until 29th Apr 2020. Both of ANN and ARIMA (0,1,0) forecasted an exponentially increasing trend for daily confirmed cases. ANN point estimates are within ARIMA interval estimates. Therefore, although ARIMA prediction is more than ANN, their differences could be neglected. The prediction of models implies that if the observed spreading pattern continues as before, the number of daily new cases would be 7872 and 9558 on 29th Apr, respectively by ANN and ARIMA models.

Regarding the goodness of fit assessment for both models; no pattern was revealed in the plots of residuals versus observations' order. Moreover, residuals seemed to be randomly scattered around zero (Fig. 3).

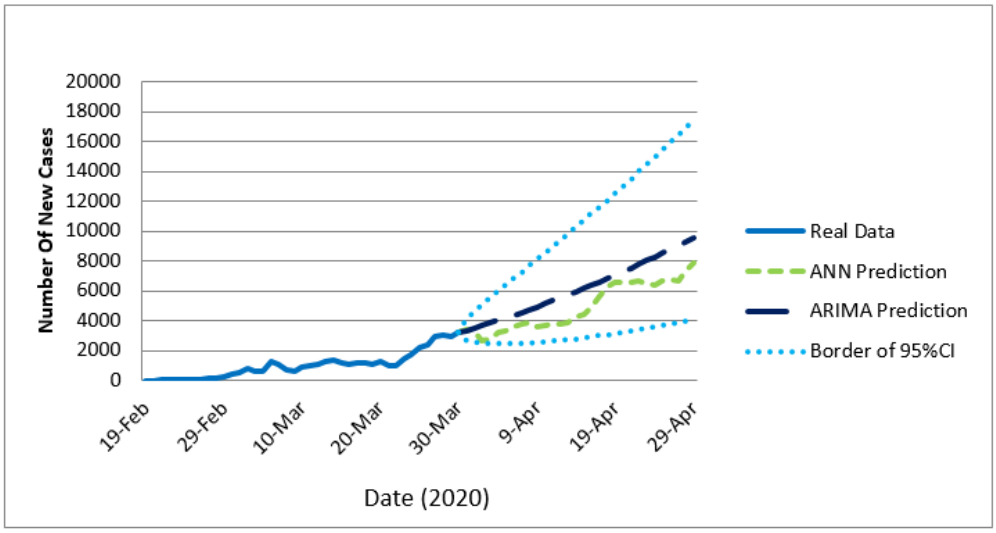


Fig. 2: Artificial Neural Netwok and ARIMA Forecasted number of new cases of COVID-19 until 29 April 2020 in Iran

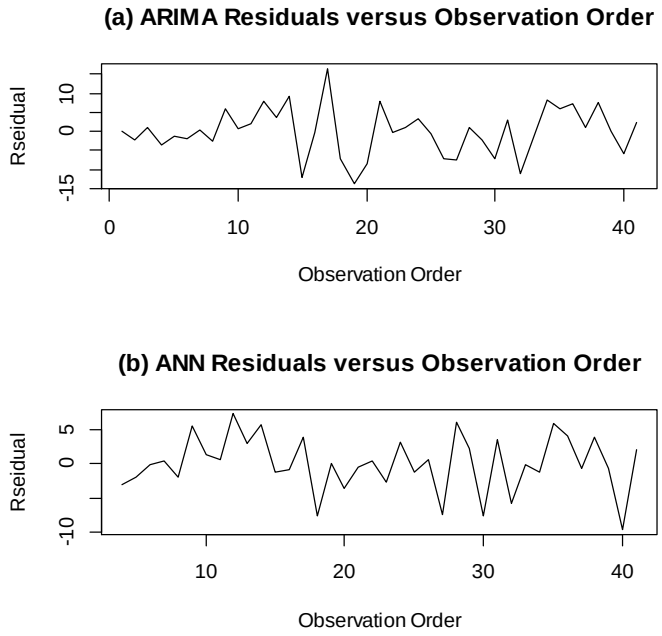


Fig. 3: Residual plots of (a) ARIMA & (b) Artificial Neural Network versus observation order

There were no spikes in Autocorrelation and Partial Autocorrelation Functions, indicating that no auto coloration remained among residuals (Fig. 4). Shapiro-Wilk tests approved normality of residuals from ARIMA and ANN algorithm (P=0.59 & P=0.23 respectively). Furthermore, NPP and Histogram of residuals did not reveal any substantial deviation from normality (Fig. 5). Box-Ljung test also affirmed the stationary of residuals from ARIMA (P=0.32) and ANN (P= 0.10). Finally, goodness of fit for ARIMA and ANN was confirmed through all residual assessments.

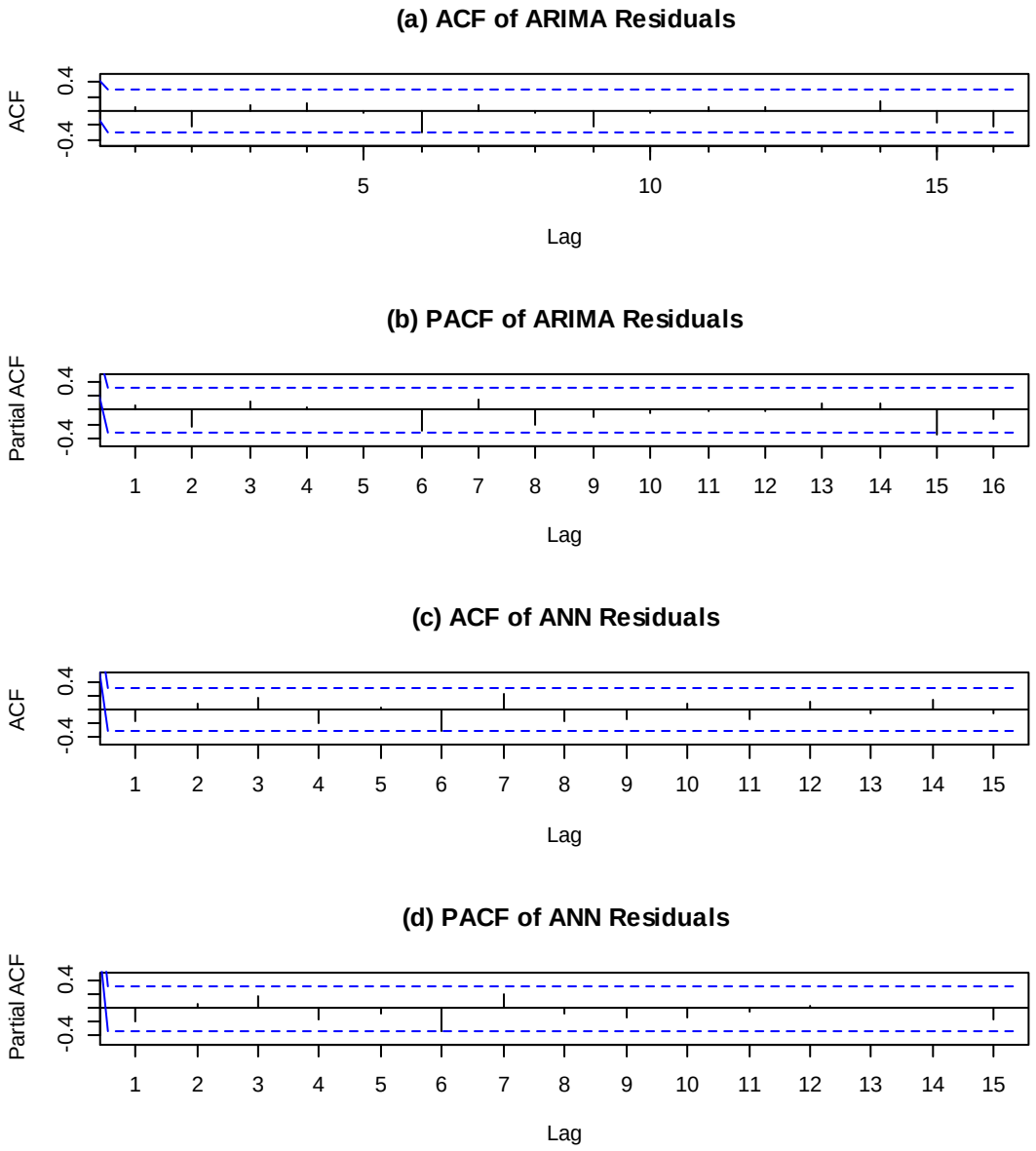


Fig. 4: Autocorrelation and Partial Autocorrelation Functions (ACF & PACF) of (a&b) ARIMA and (c&d) Artificial Neural Network residuals

The comparison of ANN and ARIMA model is embedded in Table 1, to identify the better algorithm for forecasting the new cases of COVID-19. Both MSE and MAE were less for ARIMA model; in the way that the prediction error ratio of ANN to ARIMA were 4.25 and 2.11 respectively according to MSE and MAE. Therefore the ARIMA prediction for the next coming thirty days would be more precise and more realistic. Fig. 6 also pictured the underestimation of both methods however the absolute estimation error was more for ANN algorithm. In other words, the underestimation was more for ANN in comparison to ARIMA. Therefore ANN prediction should be considered as the least ones. However even if the optimistic prediction of ANN fulfills in the future, an outbreak of COVID-19 would occur in Iran.

Table 1: Observed and Forecasted number of new cases of COVID-19 in Iran

Days (of March 2020)	***Observed***	***Forecasted***	
		ANN*	ARIMA**
25th	2206	1464	1848
26th	2389	1158	1935
27th	2926	1141	2024
28th	3076	1318	2114
29th	2901	1365	2207
30th	3189	1313	2300
MSE†		557422	2369871
MAE††		24.85	52.51

*Artificial Neural Network
**Auto Regressive Integrated Moving Average
†Mean Squared Error
††Mean Absolute Error

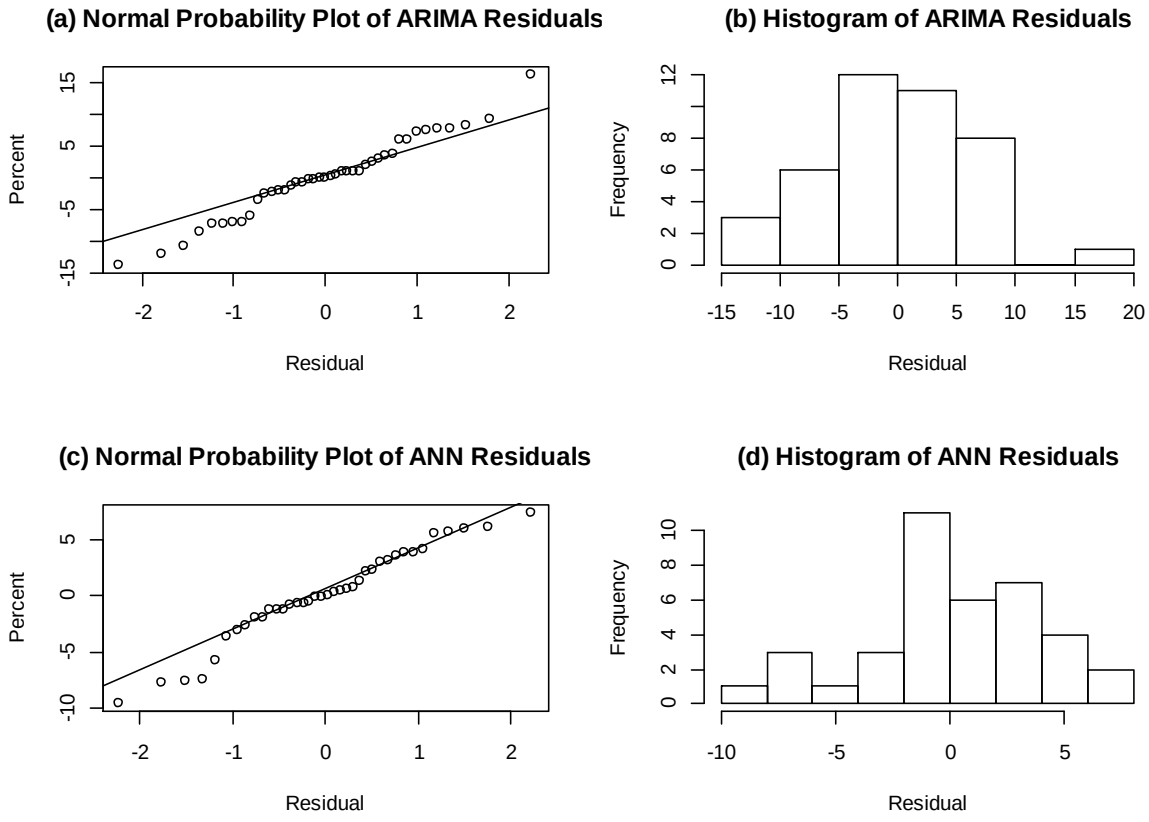


Fig. 5: Normal Probability Plot and Histogram of (a&b) ARIMA and (c&d) Artificial Neural Network residuals

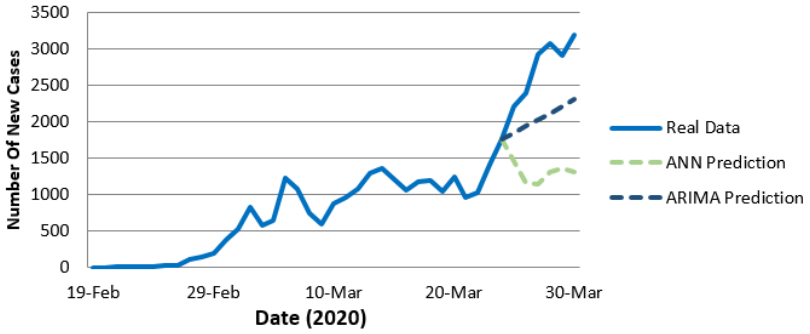


Fig. 6: Observed and Forecasted number of new cases of COVID-19 in Iran
For model comparison the observed dataset was split into two parts including: train set (from 19th Feb to 24th Mar, 2020) and test set (from 25th Mar to 31th Mar, 2020)

Discussion

In this study, we predicted the number of new cases of COVID-19 using two different models (ARIMA and ANN) and then compared them together. The results of our study showed that there would be an increasing trend in the number of new cases in the following days in Iran. The number of new cases would be 6678 and 3977 with ARIMA and ANN models 2020 respectively on 24 Apr.

Model comparison confirmed that ARIMA has more precisely forecasted future numbers. This is in line with other studies, introduced ARIMA as a useful model to predict the incidence of infectious disease (18). More accurate predictions of ARIMA in comparison to ANN has been also affirmed in applied studies previously (17).

The number of new cases will increase in Iran (12, 30-32). The number of estimated new cases are different in these studies together, which can be due to the use of different models and data at different times. Only, the ARIMA model was used and similar to one of our study models (31). The accuracy of the data is one of the issues that can affect on forecast accuracy. The lack of a sufficient number of diagnostic kits at the beginning of the epidemic and the presence asymptomatic people were the reasons that some patients were not recognized and cause underreporting and bias in forecasts.

New Year had started in Iran, and many people have gone to travel and visit relatives, while some of them may be carrying the virus, and transfer it to others. Therefore, the number of new cases may increase in the next day, and the situation in Iran will be critical. For this reason, this forecast is important for health planning, the government and health leaders that design and implement robust strategies to prevent and control the disease, because government interventions have a great impact on disease prevention.

Limitations

Our study has two major limitations. First, due to insufficient data including information about patients' demographic and their social networks, no risk factor for this disease has been evaluated. Second, few numbers of observations for this type of prediction algorithms is the major limitation in this study in the way which models might not be trained very well. However, the prediction of our study may be useful for health planning and the government.

Conclusion

Finally, if we want to have a correct forecast for the exact number of patients, we need to add data. However, we hope this paper be an alarm for health policy planners and decision-makers,

which by timely decision prepare essential materials for equipping hospitals.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors. The protocol was approved by the ethics committee of Shiraz University of Medical Sciences (No. IR.SUMS.REC.1399.066).

Acknowledgement

This study was financially supported by Shiraz University of Medical Sciences, Shiraz, Iran (Grant number 99-01-106-22291).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Tang K, Huang Y, Chen M (2020). Novel Coronavirus 2019 (Covid-19) epidemic scale estimation: topological network-based infection dynamic model. *med*R*xiv,* doi: 10.1101/2020.02.20.20023572.
2. McCall B (2020). COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. *Lancet Digit Health,* 2(4): 166-7. doi.org/10.1016/S2589-7500(20)30054-6.
3. Song PX, Wang L, Zhou Y et al (2020). An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. *med*R*xiv,* doi.org/10.1101/2020.02.29.20029421.
4. Nishiura H, Linton NM, Akhmetzhanov AR (2020). Serial interval of novel coronavirus (COVID-19) infections.*IntJInfectDis,* 93:284-6. doi: https://doi.org/10.1016/j.ijid.2020.02.060.
5. Hu Z, Ge Q, Jin L, Xiong M (2020). Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv*:2002.07112.
6. Zhao S, Gao D, Zhuang Z et al (2020). Estimating the serial interval of the novel coronavirus disease (COVID-19): A statistical analysis using the public data in Hong Kong from January 16 to February 15, 2020. *med*R*xiv,* doi:https://doi.org/10.1101/2020.02.21.20026559.
7. Zhang KK, Xie L, Lawless L, Zhou H, Gao G, Xue C (2020). Characterizing the transmission and identifying the control strategy for COVID-19 through epidemiological modeling. *med*R*xi,* doi: https://doi.org/10.1101/2020.02.24.20026773.
8. Wan H, Cui J-a, Yang G-J (2020). Risk estimation and prediction by modeling the transmission of the novel coronavirus (COVID-19) in mainland China excluding Hubei province. *med*R*xiv,* doi: https://doi.org/10.1101/2020.03.01.20029629.
9. Al-qaness MA, Ewees AA, Fan H, Abd El Aziz M (2020). Optimization Method for Forecasting Confirmed Cases of COVID-19 in China. *J Clin Med,* 9(3): 674.
10. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA (2020). The serial interval of COVID-19 from publicly reported confirmed cases. *med*R*xiv,* doi: https://doi.org/10.1101/2020.02.19.20025452.
11. Anastassopoulou C, Russo L, Tsakris A, Siettos C (2020). Data-Based Analysis, Modelling and Forecasting of the novel Coronavirus (2019-nCoV) outbreak. *med*R*xiv,* doi:https://doi.org/10.1101/2020.02.11.20022186
12. Ahmadi A, Shirani M, Rahmani F (2020). Modeling and Forecasting Trend of COVID-19 Epidemic in Iran. *med*R*xiv,* doi: https://doi.org/10.1101/2020.03.17.20037671.
13. Sun K, Chen J, Viboud C (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *Lancet Digit Health,* e201-e208.

14. Organization WH. Novel coronavirus(2019-nCoV); 2020; Available from URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.
15. Muniz-Rodriguez K, Fung IC-H, Ferdosi SR et al (2020). Transmission potential of COVID-19 in Iran. *medRxiv,* doi: https://doi.org/10.1101/2020.03.08.20030643.
16. https://www.worldometers.info/coronavirus/country/iran/
17. Hue H, Pradit S, Lim A, Goncalo C, Nitiratsuwan T (2018). Shrimp and fish catch landing trends in Songkhla lagoon, Thailand during 2003-2016. *Appl Ecol Environ Res,* 16(3): 3061-78.
18. Inoue M, Hasegawa S, Suyama A (2011). P1-177 Development and evaluation of a forecasting model for infectious diseases in Japan using time-series analysis. *J Epidemiol Community Health,* 65(1): A115-A115.
19. Bishop CM (1995). *Neural networks for pattern recognition*. Oxford university Press.
20. Zhang GP (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing,* 50: 159-75.
21. Mozhgan S, Faradmal J, Poorolajal J, Mahjub H (2017). Model-based Recursive Partitioning for Survival of Iranian Female Breast Cancer Patients: Comparing with Parametric SurvivalModels. *Iran J Public Health,* 46(1): 35-43.
22. Fattah J, Ezzine L, Aman Z, El Moussami H, Lachhab A (2018). Forecasting of demand using ARIMA model. *IJEBM,* doi: https://doi.org/10.1101/2020.03.13.20035345.
23. Medenwald D, Kuss O (2014). Mortality on match days of the German national soccer team: a time series analysis from 1995 to 2009. *J Epidemiol Community Health,* 68:869-73.
24. Ansley C, Spivey W, Wrobleski W (1977). A class of transformations for BoxJenkins seasonalmodels. *J R Stat Soc Ser C Appl Stat,* 26:173-8.
25. Günther F, Fritsch S (2010). neuralnet: Training of neural networks. *R j,* 2(1): 30-8.
26. Bellazzi R, Zupan B (2008). Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform,* 77:81-97.
27. Lee T-T, Liu C-Y, Kuo Y-H, Mills ME, Fong J-G, Hung C (2011). Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. *Int J Med Inform,* 80(2): 141-50.
28. Teshnizi SH, Ayatollahi SMT (2015). A comparison of logistic regression model and artificial neural networks in predicting of student's academic failure. *Acta Inform Med,* 23(5): 296-300.
29. Nakade M, Ojima T, Hirai H, Aida J, Hanibuchi T, Kondo K (2011). P1-259 Relations between BMI and total and cause specific mortality in Japan: ages cohort. *J Epidemiol Community Health,* 65:A138-A138.
30. Zhan C, Chi KT, Lai Z, Hao T, Su J (2020). Prediction of COVID-19 Spreading Profiles in South Korea, Italy and Iran by Data-Driven Coding. *medRxiv,* doi: https://doi.org/10.1101/2020.03.08.20032847.
31. Dehesh T, Mardani-Fard H, Dehesh P (2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. *medRxiv,* doi: https://doi.org/10.1101/2020.03.13.20035345.
32. Zheng Z, Wu K, Yao Z, Zheng J, Chen J (2020). The Prediction for Development of COVID-19 in Global Major Epidemic Areas Through Empirical Trends in China by Utilizing State Transition Matrix Model. *MedRxiv,* doi: https://doi.org/10.1101/2020.03.10.20033670.