Chaos, Solitons and Fractals 140 (2020) 110189

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using Vector Autoregressive model in Pakistan

Firdos Khan[a,*], Alia Saeed[b], Shaukat Ali[c]

[a] School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), H-12 Sector, 44000 Islamabad, Pakistan
[b] ClimatExperts, Islamabad, Pakistan
[c] Global Change Impact Studies Centre (GCISC), Ministry of Climate Change, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 28 June 2020
Accepted 4 August 2020
Available online 5 August 2020

Keyword:
Forecasting
COVID-19
Confirmed cases
Recoveries
Death
Vector Autoregressive model, Time series

ABSTRACT

COVID-19 emerged in Wuhan, China in December 2019 has now spread around the world causes damage to human life and economy. Pakistan is also severely effected by COVID-19 with 202,955 confirmed cases and total deaths of 4,118. Vector Autoregressive time series models was used to forecast new daily confirmed cases, deaths and recover cases for ten days. Our forecasted model results show maximum of 5,363/day new cases with 95% confidence interval of 3,013–8,385 on 3rd of July, 167/day deaths with 95% confidence interval of 112–233 and maximum recoveries 4,016/day with 95% confidence interval of 2,182–6,405 in the next 10 days. The findings of this research may help government and other agencies to reshape their strategies according to the forecasted situation. As the data generating process is identified in terms of time series models, then it can be updated with the arrival of new data and provide forecasted scenario in future.

1. Introduction

The world has experienced various pandemics in the past including, Spanish flue, ZIKA, Ebola, AIDS etc. which brought worst impact on human lives and economy [1–2]. The current pandemic, called COVID-19 is viral infectious disease and initially reported in Wuhan, a city of China in December last year [3]. This virus can spread in various ways, for example, by body contact, by air while infected person coughing or sneezing, through money (passing through various persons) etc. [3]. In March COVID-19 was spread in most of the countries of the world and it was considered as pandemic [3]. This pandemic has worst impact on human lives and economy due to lockdown for 2–3 months. According to worldometer's statistics about COVID-19, Pakistan is at 12th place amongst most effected countries in the world [4].

To forecast the effects on economics, finance, climatology, hydrology engineering and epidemiology etc., time series models (TSMs) play an important role [5,6]. Most of time series models need a four steps methodology starting from identification of model, estimation of unknown parameters, diagnostic checking and forecasting [7]. TSMs helps to extrapolate and predict future trends and does not simply describe existing time-dependant trends. Furthermore, to predict future disease trends, TSMs are advantageous over mechanistic models (MM) due to highly explicit epidemiological information needed to fit MMs. In contrast, TSMs need less information, consider seasonal trend and rapid fluctuations disported by diseases, unlike MMs [8]. For further details about time series modelling, we refer to [9–11].

Globally, different researchers have used time series models to model and forecast the future's scenarios of COVID-19 [12–14]. In the current pandemics of COVID-19, researchers have used different time series models to identify the data generating process and provide forecasted situations about COVID-19. [15] have used ARIMA model to model and forecast three variables of COVID-19 by utilizing the available data tell mid of April 2020. However, their results have much deviation from reality as we have observed data now. Kalman filter with ARIMA model have been used to model the situation of COVID-19 till April 30, 2020. They assumed that these variables are independent [16], however, there is dependency nature in the considered variables. To fill this gap, we proposed to model these three variables as dependant using multivariate time series model called Vector Autoregressive (VAR).

2. Data and study area

Daily data of COVID-19 new confirmed, deaths and recover cases over Pakistan from March 8 to June 27, 2020 was downloaded form World Health Organization. In Pakistan, the first

[*] Corresponding author.
E-mail address: fkyousafzai@gmail.com (F. Khan).

https://doi.org/10.1016/j.chaos.2020.110189
0960-0779/© 2020 Elsevier Ltd. All rights reserved.

Table 1
Latest statistics about COVID-19 for Pakistan and province-wise/state-wise for four variables, confirmed cases, active cases, deaths and recover cases [23]. This table has statistics for Pakistan and each state (Punjab, Sindh, Khyber Pakhtoonkhwa, Islamabad, Balochistan, Gilgit Baltistan, Azad Jammu and Kashmir).

State/provinces	Confirmed cases	Active cases	Deaths	Recover cases
Pakistan	192,970	107,760	3903	81,307
Punjab	71,191	49,327	1602	20,262
Sindh	74,070	35,480	1161	39,429
Khyber Pakhtoonkhwa	23,887	11,921	869	11,097
Islamabad	11,710	6171	115	5424
Balochistan	9817	5999	108	3710
Gilgit Baltistan	1365	347	23	995
Azad Jammu and Kashmir	930	515	25	390

COVID-19 case was recorded on March 8, 2020. Our methodology can be applied to any subregion and other regions in the world as long as you have time series data.

3. Methodology

There are various time series models which are useful for modelling and forecasting time series, including, AR, MA, ARMA, ARIMA, SARIMA, AFRIMA, and non-linear models including ARCH, GARCH etc. However, these models are useful for modelling a single time series data and we have three dependant variables. Therefore, multivariate models were used to model these three variables together. Vector Autoregressive (VAR) is a multivariate time series model and can be used to model more than one variable jointly. Suppose we have vector of time series data Y_t, then a VAR model with k variables and p lags can be expresses mathematically in Eq. (1).

$$\boldsymbol{Y_t} = \boldsymbol{B}_0 + \boldsymbol{BY}_{t-1} + \boldsymbol{B}_2\boldsymbol{Y}_{t-2} + \ldots + \boldsymbol{B_p}\boldsymbol{Y}_{t-p} + \boldsymbol{\epsilon_t} \tag{1}$$

Where in Eq. (1), $\boldsymbol{Y_t}$,$\boldsymbol{B}_0$ and $\boldsymbol{\epsilon_t}$ are $k \times 1$column vector and $\boldsymbol{B}_0$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, ..., $\boldsymbol{B_p}$ are $k \times k$ matrices of coefficients. The simplest VAR model for three variables is with lag $p = 1$ and can be expressed as in Eq. (2)

$$\begin{bmatrix} y_{1(t)} \\ y_{2(t)} \\ y_{3(t)} \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \\ b_{30} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \begin{bmatrix} y_{1(t-1)} \\ y_{2(t-1)} \\ y_{3(t-1)} \end{bmatrix} + \begin{bmatrix} \epsilon_{1(t)} \\ \epsilon_{2(t)} \\ \epsilon_{3(t)} \end{bmatrix} \tag{2}$$

A four-step methodology was adopted to accomplish this study. The first step was identification of the model or the lag-length selection. For this purpose, model selection criteria including Akaike's Information Criterion (AIC) [17], Hannan–Quinn criterion [18], Schwarz criterion [19] and Final Prediction Error (FPE) [20] have been used. Second step is the estimation of unknown parameters of the identified model and ordinary least square (OLS) method was used for this purpose. Diagnostic checking is an important step in time series modelling which is comprises analysis of the residuals of the fitted model graphically as well as by statistical tests. Graphical investigation includes histogram, ACF and PACF of residuals while statistical tests include serial correlation and normality tests. Finally, forecasting of the modelled time series can be made which is the ultimate objective of time series modelling. Once the data generating process is investigated and identified in the shape of time series models, then it can be used for h-steps ahead forecasting. The forecast can provide guidelines for actions and policy making as it provides future scenarios. For analysis, we used R software with packages vars [21] and mFilter [22].

4. Results and discussion

The latest statistics about pandemic COVID-19 about Pakistan and state-wise are given in Table 1. There is rapid increase in new cases of Covid-19 after the Ediul-fitar festival. The confirmed cases jumped from 55,000 to 202,955 in one month when the government released the lockdown and the people did not care about COVID-19 SOPs. Currently, Pakistan has 106,213 active, 4118 deaths and 92,624 recover cases from the pandemics of COVID-19 is shown in Table 1 and Fig. 1 with clear increasing trends. VAR model was used for modelling and forecasting by implementing four step time series modelling procedures. In the first step model lags-length was identified by using model selection criteria, where two criteria (AIC and FPE) suggest 5 lags while HQ and SC are in favour of 2 and 1 lags, respectively. Consequently, we used 5 lags in VAR model for further analysis. Unknown parameters of identified model were estimated by using ordinary lest square (OLS) method of estimation. After parameters estimation, it is important to make a diagnostic check which includes graphical investigation and statistical testing of the residuals of the fitted model. Graphical investigation of residual's was carried out by plotting, histogram, ACF and PACF of fitted model. Serial correlation test and normality test were performed and found that the estimated model is fine and qualify diagnostic checks. Furthermore, the estimated model was investigated for ARCH (ARCH = Autoregressive Heteroscedasticity) error and rejected the hypothesis of arch error in the fitted model. After all these steps, the model is ready for forecasting. In this study, we performed a 10-days ahead forecasting, however, it depends on the purpose and objectives of the study to have h-step ahead forecasting.

The 10-days ahead forecasted result of daily new cases is presented in Fig. 2 where it can be seen that the daily new cases are increasing. The maximum number of cases forecasted in 10 days is 5363 with 95% confidence intervals of 3013 and 8385. The minimum number of daily new cases forecasted are 2861 with 95% confidence intervals of 1494 and 4665. It is worth mentioning that the forecasted daily new cases follow the same pattern and variability of observed data. Fig. 3 shows the observed, forecasted and 95% confidence intervals for daily deaths due to COVID-19 in Pakistan. The results show that the daily deaths follow increasing trend like that of observed data. Maximum daily deaths forecast reached to 167 with 95% confidence intervals of 112 and 233. The minimum number of daily deaths forecasted are 93 with 95% confidence intervals of 54 and 142. About daily recoveries, the forecasted results follow the same pattern and variability. The forecasted results show that maximum recoveries can reach to 4016 with 95% confidence intervals of 2182 and 6405 in the next 10 days. The minimum daily recover cases can be 967 with 95% confidence intervals of 225and 2227 in the upcoming 10 days. The

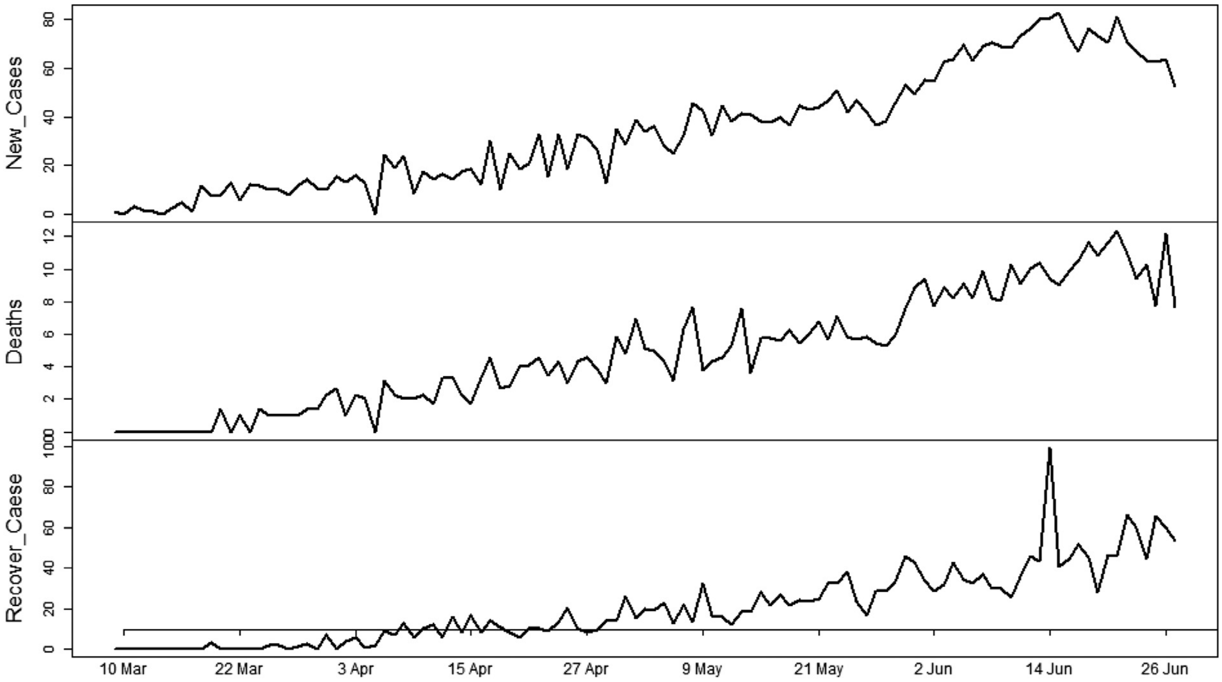

Fig. 1. Daily new cases, deaths and recover cases after the start of pandemic COVID-19 in Pakistan for the duration of March 8-June 27, 2020. On y-axis and x-axis, number of cases for each variable and time are mentioned, respectively.

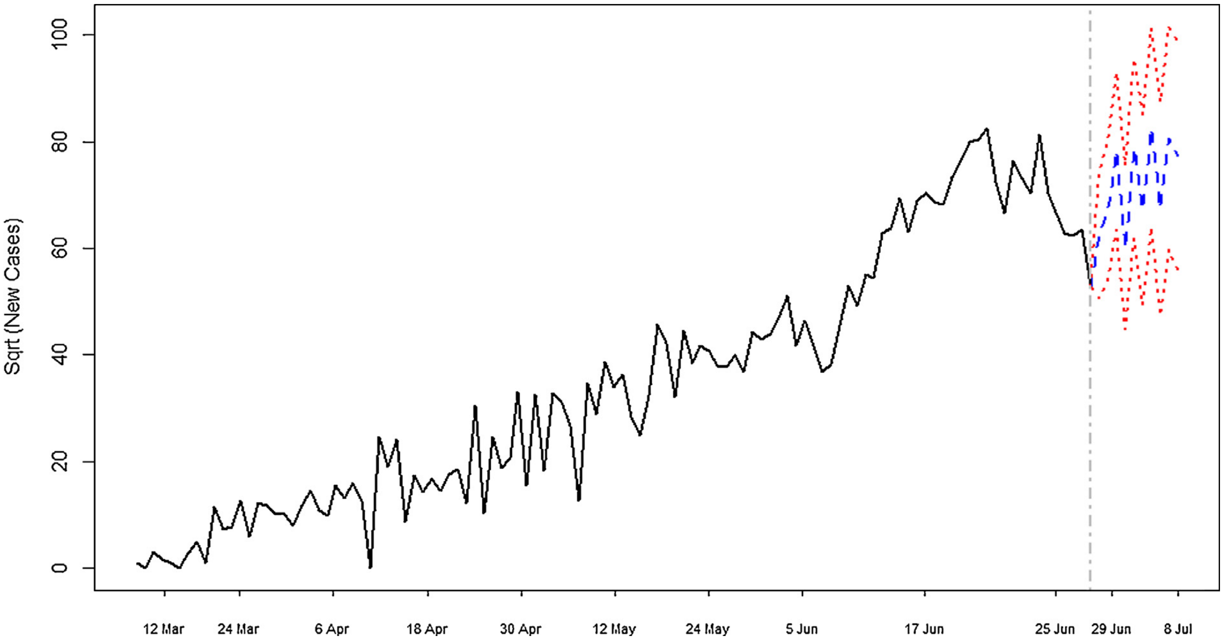

Fig. 2. 10-days ahead forecast of daily neas cases with 95% confidence intervals in Pakistan. The black, blue, red colors show observed, forecasted and 95% confidence intervals for daily new cases, respectively. On y-axis and x-axis, number of new cases and time are mentioned, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

results of [15] suggested an increase of 2.7 times in new cases, eightfold increase in recoveries and maximum forecasted deaths were 500 in the end of May 2020 in Pakistan, however, in reality these cases did not hit such high values for both variables. There is an increasing trend forecasted for 5 days in new cases, deaths and recoveries during May 1–5, 2020 where maximum numbers were 15,652, 516 and 6342 for total active cases, deaths and recover cases, respectively, in Pakistan [16]. In contrast, our forecasted results have reasonable values in the next ten days about COVID-19 for Pakistan where maximum forecasted values are 5363, 167 and 4016 for new cases, deaths and recover cases, respectively. The possible reasons may be less data (till the end of April

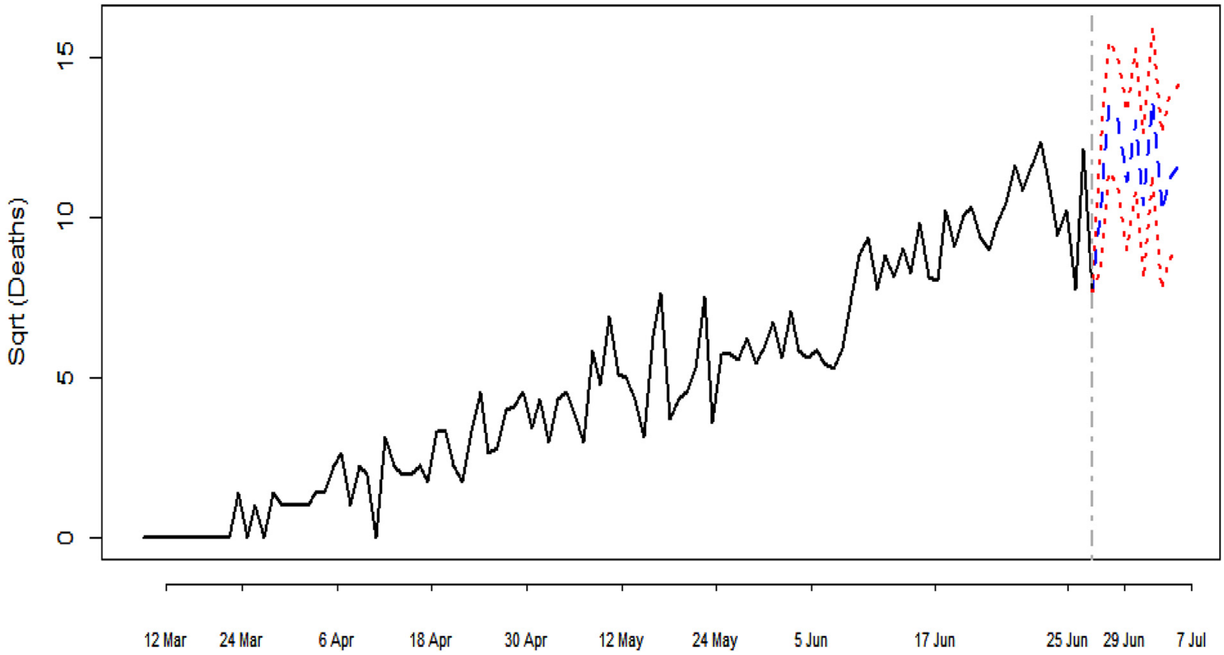

Fig. 3. 10-days ahead forecast of daily deaths with 95% confidence intervals in Pakistan. The black, blue, red colors show observed, forecasted and 95% confidence intervals for daily deaths, respectively. On y-axis and x-axis, number of deaths and time are mentioned, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

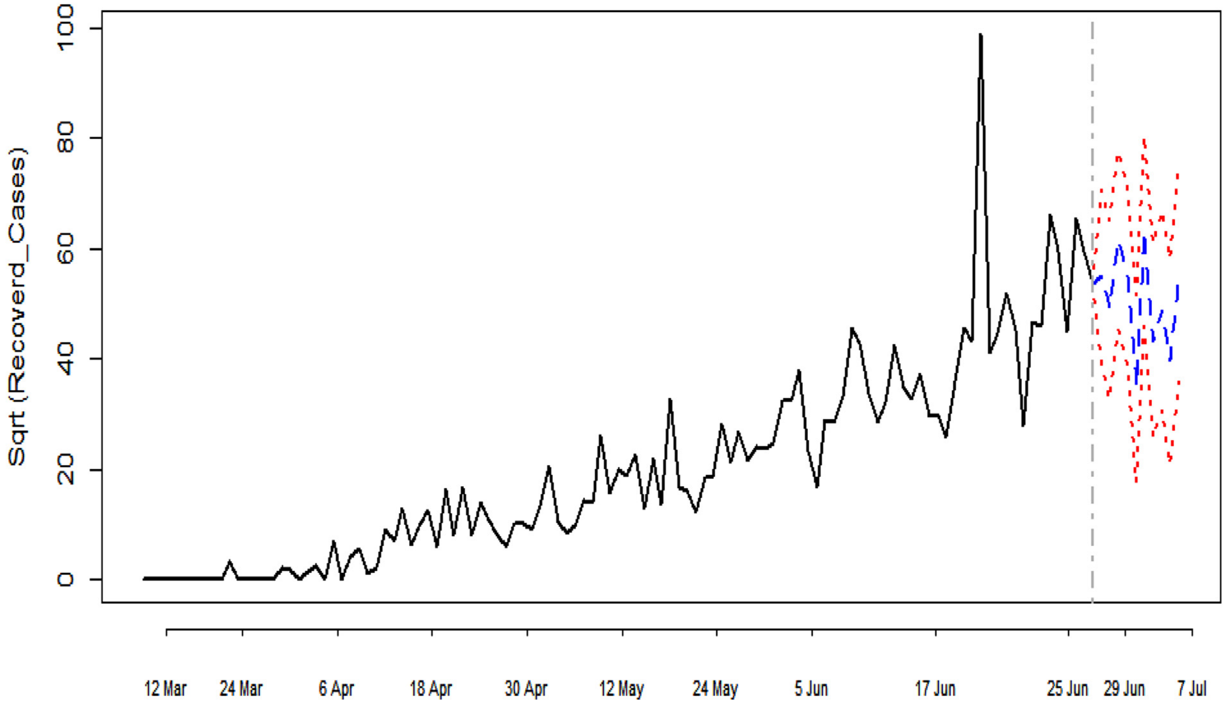

Fig. 4. 10-days ahead forecast of daily recover cases with 95% confidence intervals in Pakistan. The black, blue, red colors show observed, forecasted and 95% confidence intervals for daily recover cases, respectively. On y-axis and x-axis, number of recover cases and time are mentioned, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2020) and univariate TSMs in [15,16]. A deep learning approach, long short-term memory (LSTM) was used for forecasting the possible end of COVID-19 in Canada and concluded that the possible end will be around June 2020 [13]. Autoregressive (AR) time series model with two–piece scale mixture normal (TP–SMN) distributions was used to model and forecast confirmed and recovered cases of COVID-19 by utilizing data from 22 January to 30 April 2020 in Iran [12]. Gane Expression Programming (GEP) was used to forecast confirmed and deaths cases in various states of India [24]. Their results have increasing trends and therefore, suggested strict lockdown and social distancing to keep virus in control.

5. Summary and conclusion

This study is an attempt to provide future's scenario about COVID-19 in Pakistan. In the first step, data generating process was identified in terms of time series models. In this study a multivariate time series model known as Vector Autoregressive (VAR) was identified with 5 lags using various model selection criteria. The unknown parameters of identified model were estimated by using ordinary least square method of estimation. Diagnostic checks of the estimated model were carried out by using graphical investigation and using statistical test to the residuals of the fitted model. A 10-days ahead forecast was performed for daily new cases, daily deaths and daily recover cases with 95% confidence intervals. The forecasted results revealed that maximum and minimum number of daily new cases may be 5363 and 2.861, respectively, in the upcoming 10 days. The 10-days ahead forecasted results show that minimum and maximum number of daily deaths will be 93 and 167, respectively. Minimum and maximum forecasted results of daily new recoveries can be 967 and 4016, respectively, in the next 10 days. The results of this study may be helpful for policy makers and other stakeholders in health and other departments. One limitation of this study is availability of less data about COVID-19 as with more data, the parameters of the model are stable robust and consequently more useful for forecasting.

Declaration of Competing Interest

The authors declare no conflict of interest.

CRediT authorship contribution statement

Firdos Khan: Conceptualization, Methodology, Formal analysis, Writing - review & editing, Validation, Visualization. **Alia Saeed:** Data curation, Software. **Shaukat Ali:** Writing - review & editing.

Acknowledgement

The authors are thankful World Health Organization and National Institute of Health of Pakistan for providing data about COVID-19. This research has no funding resources.

References

[1] Bloom DA, Cadarette D, Sevilla JP. Epidemics and economics. International Monetary Fund (IMF). Finance Dev 2018;55(2).
[2] Madhav N, Oppenheim B, Gallivan M, et al. Pandemics: risks, Impacts, and Mitigation. Disease control priorities: improving health and reducing poverty. Jamison DT, Gelband H, Horton S, et al., editors. 3rd edition, Washington (DC): The International Bank for Reconstruction and Development/The World Bank; 2017 Nov 27. Chapter 17. doi:10.1596/978-1-4648-0527-1_ch17.
[3] World Health Organization (WHO). URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (Accessed on June 26, 2020).
[4] Worldometers' statistics about Corona virus pandemics (COVID-19) https://www.worldometers.info/coronavirus/ (Accessed on June 25, 2020).
[5] Mathevet T, Lepiller M, Mangin A. Application of time series analyses to the hydrological functioning of an Alpine karstic system: the case of Bange-L'Eua-Morte'. Hydrol Earth Syst Sci 2004;8(6):1051–64.
[6] Khan F, Pilz J. 'Modelling and sensitivity analysis of river flow in the Upper Indus Basin, Pakistan. Int J Water 2018;12(1):1–21.
[7] Box GEP, Jenkins S. Time series analysis, forecasting and control. San Francisco: Holden-Day; 1970.
[8] Zhang X, Liu Y, Yang M, Zhang T, Young AA, Li X. Comparative study of four time series methods in forecasting typhoid fever incidence in China. PLoS ONE 2013:e63116. doi:10.1371/journal.pone.0063116.
[9] Barber D, Cemgil AT, Chiappa S. Bayesian time series models. Cambridge, UK: Cambridge University Press; 2011.
[10] Enders W. Applied econometric time series. 3rd editor. New York, USA: John Wiley and Sons; 2010.
[11] Douc R, Moulines E, Stoffer DS. Nonlinear time series: theory, methods, and application with R examples. USA: CRC Press, Taylor & Francis Group; 2014.
[12] Maleki M, Mahmoudi MR, Wraith D, Pho KH. Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel Med Infect Dis. 2020. https://doi.org/10.1016/j.tmaid.2020.101742.
[13] Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fract 2020;135:109864 2020 Jun. doi:10.1016/j.chaos.2020.109864.
[14] Yonar H, Yonar A, Tekindal MA, Tekindal M. Modeling and forecasting for the number of cases of the COVID-19 pandemic with the curve estimation models,the box-jenkins and exponential smoothing methods. Eur J Med Oncol 2020;4(2):160–5. doi:10.14744/ejmo.2020.28273EJMO.
[15] Yousaf M, Zahir S, Riaz M, Hussain SM, Shah K. Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan. Chaos, Solitons Fract 2020;138:109926. https://doi.org/10.1016/j.chaos.2020.109926.
[16] Aslam M. Using the kalman filter with Arima for the COVID-19 pandemic dataset of Pakistan. Data Brief 2020;31:105854. https://doi.org/10.1016/j.dib.2020.105854.
[17] Akaike H. Statistical predictor identification. Ann Inst Statist Math 1970;22:203–17.
[18] Hannan EJ, Quinn BG. The determination of the order of autoregression. J R Statist SocB 1979;41:190–5.
[19] Schwarz G. Estimating the dimension of a model. Ann Statist 1978;6:461–4.
[20] Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. 2nd International symposium on information theory. Budapest, Hungry: Akademia Kiado; 1973. p. 267–81.
[21] Pfaff B. Package vars 2018 https://cran.r-project.org/web/packages/vars/vars.pdf. Accessed on 27 June 2020.
[22] Baleilar M. Package mFilter 2019 https://cran.r-project.org/web/packages/mFilter/mFilter.pdf. Accessed on 27 June 2020.
[23] Government of Pakistan official portal for COVID-19. URL: http://covid.gov.pk/stats/pakistan (accessed on June 27, 2020).
[24] Salgotra R, Gandomi M, Gandomi AH. Time series analysis and forecast of the COVID-19 Pandemic in India using genetic programming. Chaos Solitons Fractals 2020. doi:10.1016/j.chaos.2020.109945.