Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Regression modeling and correlation analysis spread of COVID-19 data for Pakistan

Dure Jabeen [1,*], Ingila Rahim [1], Rumaisa Iftikhar [1], M. Rafiullah [2], M. Rashid Kamal Ansari [1]

[1]Department of Electronics and Mathematics, Sir Syed University of Engineering and Technology, Karachi, Pakistan
[2]Department of Mathematics, COMSATS University Islamabad, Lahore, Pakistan

ARTICLE INFO

Article history:
Received 25 September 2021
Received in revised form
3 January 2022
Accepted 3 January 2022

Keywords:
COVID-19
ARIMA
Correlation
Forecast
Confirmed cases
Death cases
Recovery cases

ABSTRACT

This study presents a mathematical analysis of the coronavirus spread in Pakistan by analyzing the (COVID-19) situation in six provinces, including Gilgit Baltistan, Azad Jammu Kashmir and federal capital (seven zones) individually. The influence of each province and the Federal Capital territory is then observed over the other territories. By subdividing the associated data into confirmed cases, death cases, and recovery cases, the dependence of the (COVID-19) situation from one province to the other provinces is investigated. Since the worsening circumstance in the neighboring countries were considered as a catalyst to initiate the outburst in Pakistan, it seemed necessary to have an understanding of the situation in neighboring countries, particularly, Iran, India, and Bangladesh. Exploratory data analysis is utilized to understand the behavior of confirmed cases, death cases, and recovery cases data of (COVID-19) in Pakistan. Also, an understanding of the pandemic spread during different waves of (COVID-19) is obtained. Depending on the individual situation in each of the provinces, it is expected to have a different ARIMA model in each case. Hunt for the most suitable ARIMA models is an essential part of this study. The time-series data forecasts by processing the most suitable ARIMA models to observe the influence of one territory over the other. Moreover, forecasting for the month of August 2021 is performed and a possible correlation with actual data is determined. Linear, multiple regression, and exponential models have been applied and the best-fitted model is acquired. The information obtained from such analysis can be employed to vary possible parameters and variables in the system to achieve optimal performance.

1. Introduction

The novel coronavirus has corroborated itself to be one of the most fatal pandemics of this century. Globally, the earliest cases of coronavirus were noticed in December 2019 in the city of Wuhan, the capital of Hubei province, China (WHO, 2020), and the situation was declared as pandemic on March 11th, 2020. This virus typically attacks a person's respiratory system and the disease commonly known as COVID-19. It has an aptitude to transmit itself at an alarmingly quick pace, thus dispersal of the virus across the globe became a matter of a few months (Tomar and Gupta, 2020). To counter the spread of the disease, protective measures, enhancement of medical facilities and research and production of a possible vaccine overburdened the global financial budgets. Strict lockdowns caused huge losses to world economies. Some facts about the situation are summarized as follows:

1. An unceasing enhancement in expenses to support medical necessities.
2. A continued effort to keep the infection restricted in a specific domain.
3. To assure unfailing precautionary measures.
4. To keep the general public aware of the severity of the situation to avoid the appearance of any panic.
5. To overcome the indirect consequences resulting in cessation of businesses activities and steep rise in the unemployment rate (Kawohl and Nordt, 2020).
6. To extend testing and vaccination services to the general public protracted and exceptional test of

* Corresponding Author.
Email Address: durejabeen@ssuet.edu.pk (D. Jabeen)
https://doi.org/10.21833/ijaas.2022.03.009
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-6743-2911

public mental health, with foremost inferences of suicide (Gunnell et al., 2020).

The key corrective measure for COVID-19 is the maintenance of social distancing. This strategy was found successful to decelerate the outburst of the pandemic in a few countries (Anderson et al., 2020). The Transmission Rate of COVID-19, denoted by Ro, by major research studies is estimated to be somewhere between (1.5 to 3.5) (Imai et al., 2020; Read et al., 2021) which is quite high.

Mathematical models help in understanding the transmission, spread, recovery, and death due to the virus, hence playing a vital part in policy making and constructing important decisions in the public health sector through evidence-derived statistics (Khajanchi et al., 2018). Across the globe, scientists have been attempting to produce an effective model to depict the behavior of COVID-19. In Jung et al. (2020), the researchers offered a model forecasting the rate of mortality from COVID-19. One such study (Jung et al., 2020) predicted the rate of mortality to be 5.1% with a reproduction rate of 2.1 while the other (Jung et al., 2020) estimated it to be 8.4% with 3.2 reproductions. The estimation results hinted at the danger of an upcoming pandemic. Susceptible-Infected-Removed (SIR) model was one of the earliest established models, for the appropriate forecast of the coronavirus outburst in China (Zhong et al., 2020). The SIR model was also utilized to predict the COVID-19 outbreak in Iran; the parameters were estimated via Generalized Additive Model (GAM) models (Zareie et al., 2020). Delivering actual information associated with COVID-19 with a statistical analysis of figures, Susceptible-Exposed-Infectious-Recovered (SEIR) modeling was utilized for estimating on daily basis by creating micro-services to extract data via various sources (Hamzah et al., 2020). SEIR modeling was also utilized with the data provided for countries South Korea, Italy, and Iran for forecasting the COVID-19 dispersion profile (Zhan et al., 2020). NNETAR, ARIMA, Hybrid, Holt-Winter, BSTS, TBATS, Prophet, MLP, and ELM network models were compared with Iranian data to determine the model with the lowest error in forecasting (Talkhi et al., 2021). Recurrent neural network (RNN) grounded variants of long short-term memory (LSTM) were applied to construct models to forecast future timelines for India and USA (Shastri et al., 2020). LSTM method along with classical curve fitting was implemented to predict the number of new patients in India (WHO, 2021). Adaptive Neuro-Fuzzy Inference System (ANFIS) estimated the number of forthcoming cases for the next ten days in China with the help of data provided by WHO. Time-series analysis (Li, 2020; Guo et al., 2020) is an effective means for future estimation. It aids in generating a mathematical model with respect to the consistency and trend of the previously observed statistics against time. There are multiple time series evaluation models that are tested to provide fruitful results for the monitoring of virus control. Auto-Regressive Integrated Moving Average (ARIMA) model is implemented in our research. In the cases of infectious disease, this technique is generally implemented for time series prediction. This model, originally designed for economic applications, has proven its worth in the medical field too. The principle of this estimation consists of filtering of the high-frequency noise in the statistics, sensing the local tendencies, grounded on linear dependence, and predicting the future trends (Kane et al., 2014). In Wang et al. (2018), authors utilized the ARIMA model and grey model GM (1,1) to forecast the number of cases of hepatitis B in China with the help of data acquired over seven years, ARIMA provided a better estimation performance as compared to the grey model. The Coronavirus disease (COVID-19) broke out in Wuhan in 2019. By early 2020, it was declared as a pandemic. Pakistan reported its first case from the capital city, Islamabad in February 2020. Pilgrims from Iran were considered as the primary source for carrying this contagious disease in Pakistan at that time. This study analyses the (COVID-19) spread phenomenon in Pakistan from February 2020 to August 2021, when the third wave had much progressed and the fourth wave of (COVID-19) was preparing to take off. Following (WHO, 2021) ARIMA models are used to study the available data for each province. Time-series data analysis has been performed for the respective data using different regression models and exploratory analysis determined to understand the virus spread, data correlation and visualization has been discussed in the following five sections namely, introduction, Pakistan COVID data, research methodology, results discussion, and conclusion.

2. Pakistan COVID-19 data

COVID-19 data of Pakistan has been taken from the Ministry of National Health Services Regulations and Coordination. This statistic is divided into 7 provinces namely Sindh, Punjab, Balochistan, Khyber Pakhtunkhwa (KPK), Islamabad, Azad Jammu and Kashmir (AJK), and Gilgit-Baltistan (GB). Islamabad, Azad Jammu and Kashmir, and Gilgit-Baltistan are considered as provinces of Pakistan for COVID-19 propagation. The data set is further segregated in confirmed cases (CC) 1159,427, death cases (DC) 25,599 and recovery cases (RC) 1042,734; it has been observed from March 2020 to date. Pakistan COVID-19 data has been processed and analyzed for March 2020 to August 2021, the future forecasting has been presented for the month of August 2021 and its correlation is performed with actual observed data.

3. Research methodology

ARIMA: The exploratory data analysis is used to understand its frequency distribution over the entire region. This data can be considered as discrete-time series (signal for processing and analysis) for analytical analysis. The total number of cases

statistics has been processed and observed for mean value, standard deviation, skewness, and kurtosis. Further Box-Jenkins model has been applied which is known as Autoregressive Integrated Moving Average (ARIMA) model. It has been considered a powerful method to analyze forecasts and predict the future conduct of the time series. This model is based on its own past value and error terms that are based on its own lags and forecast error of lagged. It has been applied to stationary data. ARIMA is characterized by three order terms (p, d, q) with AR(p) and MA(q). The general term of ARIMA is defined as below:

$$Y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \ldots + \varphi_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

$$y_t = \frac{Constant\ term + Linear\ Combination\ of\ Y\ lag\ upto\ p\ value + Linear\ combination\ of\ lags\ for\ forecast\ errors\ upto\ q\ lags}{}$$

where, Y_t is the predicted value which is defined in words as given above. It has been estimated on autoregressive (AR), integrated (I), and moving average (MA) for different orders. The statistical data of COVID-19 is non-stationary, and it is converted into stationary data by applying different tests like unit root (augmented dickey fuller) with the different operations (Sharma et al., 2020). The model identification has been performed with correlation. Correlation function has been found on stationary data to pick the AR(p) and MA(q) order terms with partial correlation (PACF) and autocorrelation function (ACF). The models have been estimated with different orders of *(p, d, q)*. The best fit values are selected for all provinces by Akaike's information criterion (AIC) on an individual basis. The best model diagnostic checking has been observed with Q-statistic and it was noted that all roots of the best fit model were lying inside the unit circle. The selected models have been used to forecast the future values of the process. Moreover, the correlation between July-August 2021 real data and forecasted data has been performed to test the selected model performance.

3.1. Regression model

The regression analysis has been performed because it provides tremendous flexibility in different circumstances. It is applied here to understand the relationship between the independent and dependent variables on the basis of their coefficients. The confirmed, recovered and death cases are considered as dependent variables whereas months, average month temperature, and humidity are presented as independent variables. The following method and analysis have been performed: Linear regression for total COVID 19 cases, multiple regressions have been applied on confirmed, recovered and death cases with monthly growth, average temperature, and humidity; and exponential growth model (EGM) with respect to monthly growth.

3.2. Linear regression model (LRM)

It is described as below with the parameters $\alpha_\circ$, $\beta_\circ$ are the slope/gradient and intercept respectively. $\alpha_\circ$, $\beta_\circ$ represents the growth rate of the spread COVID-19 on daily basis. P_n and Y_n is the daily growth of confirmed, recovered, death cases and represents the day or a month under consideration respectively. The error term can be calculated with given following equation.

$$P_n = \alpha_o + \beta_o Y_n \quad (2)$$

$$P_n = Intercept + Total(C, or\ R, or\ D) coefficients Y_n$$

$$e_n = P_n - \alpha_o - \beta_o Y_n \quad (3)$$

3.3. Multiple regressions (MR)

It is the relationship between two or more independent variables and modeled with corresponding coefficients such as:

$$P_n = a_o + b_1 Y_1 + b_2 Y_2 + b_3 Y_3 + \ldots \quad (4)$$

3.4. Exponential growth model (EGM)

It is the relationship between the months and increasing rate of affected cases as the cumulative sum on the monthly basis, EGM equation is presented here:

$$P_n = a_o e^{b_o Y_n} \quad (5)$$

4. Results and discussion

COVID-19 data of Pakistan is selected and divided into 7 provinces as discussed in section "Pakistan COVID-19 data."

4.1. Tail analysis

The Exploratory data analysis has been applied with data of CC, DC, and RC to understand the behavior of COVID-19 statistics. Table 1 shows the total number of cases, mean, median, standard deviation, skewness, and kurtosis of time series with a random process. Data frequency of distribution has been observed by the skewness and kurtosis. The random process statistics of all provinces are positively skewed for CC, DC, and RC. And others have lower frequency distribution than normal. Figs. 1-3 shows the CC, DC, and RC data of COVID-19 in all provinces from March 2020 to August 2021.

The correlation of COVID-19 spread among all provinces has been calculated and presented in Tables 2-4 with CC, DC, and RC respectively. The correlation among Sindh, Punjab, Baluchistan, KPK, and Islamabad is higher as these are population-dense areas. The correlation between AJK and GB with other provinces is comparatively weaker, being the less population density area. Similarly, the RC

among Sindh, Punjab, Baluchistan, KPK, and Islamabad is higher than AJK and GB. It is because they have better medical facilities than AJK and GB.

Table 1: Exploratory data analysis of CC DC and RC

			Confirmed Cases		
Area	Mean	Median	SD	Skewness	Kurtosis
Sindh	720.5	521	646.2	0.98	0.41
Punjab	467.7	329	498.7	2.01	4.72
Balochistan	56.9	33	69.1	2.9	11.32
KPK	199.44	167	175.02	1.08	1.18
Islamabad	122.59	76	139.05	1.58	2.35
AJK	26.72	17	29.7	1.61	2.73
GB	14.57	12	12.117	0.91	0.27
			death Cases		
Sindh	11.71	8	11.943	1.34	1.41
Punjab	13.356	7	15.713	1.62	2.78
Balochistan	0.576	0	1.126	3.81	24.31
KPK	5.641	4	5.884	1.19	1.39
Islamabad	1.41	1	1.84	1.48	2.04
AJK	0.79	0	1.195	1.91	3.86
GB	0.31	0	0.66	2.40	6.05
			Recovery Cases		
Sindh	664.8	361	1084.8	6.8	72.04
Punjab	423.8	61	1320	6.6	50.61
Balochistan	54.85	27	72.67	2.09	4.15
KPK	184	85	270.1	3.97	23.56
Islamabad	117.33	47	161.2	3.26	21.91
AJK	24.83	10	34.7	2.31	7.75
GB	14.226	11	14.95	1.38	2.39

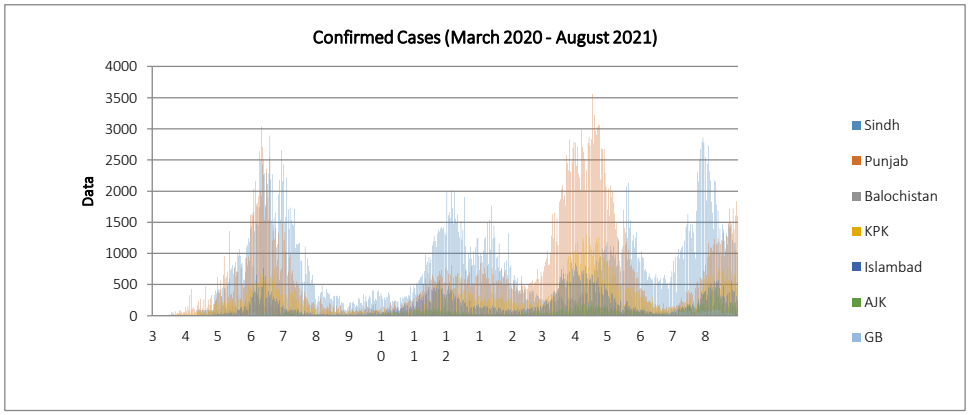


Fig. 1: Confirmed cases for all provinces

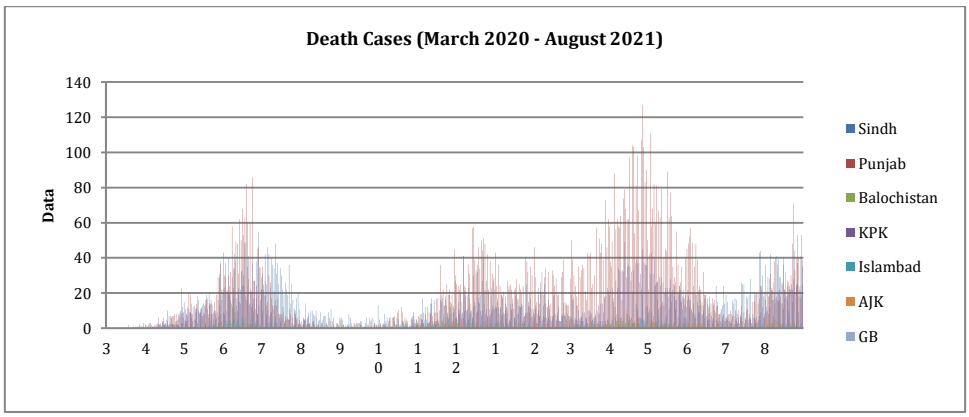


Fig. 2: Death cases for all provinces

Table 2: Correlation matrix for CC of all provinces of Pakistan March 2020 to August 2021

	Sindh	Punjab	Balochistan	KPK	Islamabad	AJK	GB
Sindh	1	0.35	0.45	0.41	0.30	0.17	0.13
Punjab		1	0.42	0.91	0.81	0.68	-0.08
Balochistan			1	0.36	0.26	0.08	0.41
KPK				1	0.82	0.73	-0.08
Islamabad					1	0.78	-0.07
AJK						1	-0.12
GB							1

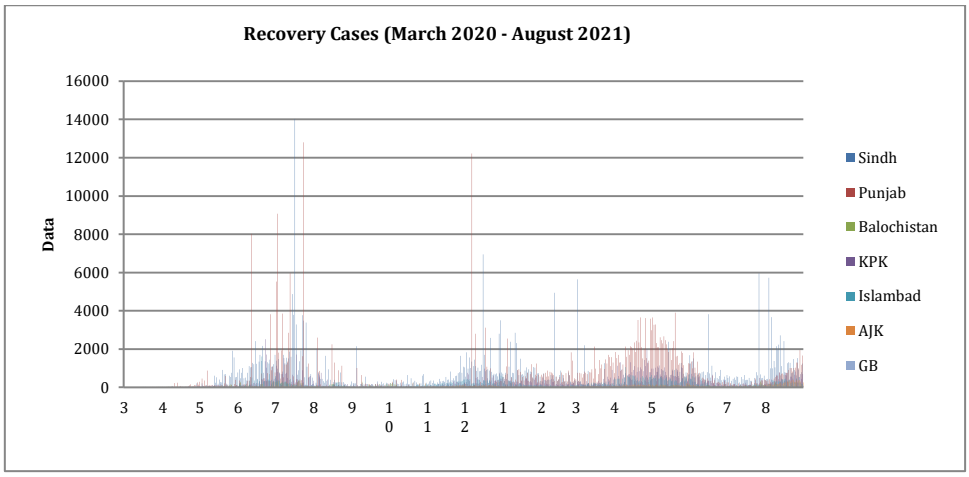


Fig. 3: Recovery cases for all provinces

Table 3: Correlation matrix for DC of all provinces of Pakistan March 2020 to August 2021

	Sindh	Punjab	Balochistan	KPK	Islamabad	AJK	GB
Sindh	1	0.26	0.33	0.21	0.24	0.04	0.19
Punjab		1	0.26	0.76	0.47	0.48	-0.08
Balochistan			1	0.31	0.26	0.07	-0.01
KPK				1	0.44	0.49	-0.07
Islamabad					1	0.41	0.41
AJK						1	-0.01
GB							1

Table 4: Correlation matrix for RC of all provinces of Pakistan March 2020 to August 2021

	Sindh	Punjab	Balochistan	KPK	Islamabad	AJK	GB
Sindh	1	0.14	0.24	0.20	0.16	0.07	0.04
Punjab		1	0.27	0.38	0.41	0.33	-0.03
Balochistan			1	0.27	0.21	0.09	0.25
KPK				1	0.57	0.54	-0.01
Islamabad					1	0.63	-0.05
AJK						1	-0.06
GB							1

4.2. COVID-19 forecasts for Pakistan

Further, the Autoregressive Integrated Moving Average Model (ARIMA) has been constructed for all provinces' CC, DC, and RC with recorded data. The first two orders have been selected to find ARIMA *(p, d, q)* model. The best models have been selected by Akaike's information criterion (AIC) shown in Table 5. The equation of Best ARIMA models for CC, DC, and RC are shown in Tables 6-8 respectively. The correlation between August recorded data and forecasted data has been shown in Table 9, which is also presented in Fig. 4. It has been observed that during the second wave, numbers of cases have increased in highly populated areas like Sindh, Punjab, Balochistan, Islamabad, and KPK. The death rate is minimum due to COVID-19 smart lockdown in Pakistan. It has been observed that the numbers of deaths are minimum in the lowest populated cities like AJK and GB.

Table 5: Best of ARIMA model selection by AIC

	Confirmed Cases		Death Cases		Recovery Cases	
Area	ARIMA(p,d,q)	AIC	ARIMA(p,d,q)	AIC	ARIMA(p,d,q)	AIC
Sindh	(1,1,1)	13.7	(1,1,1)	6.76	(1,1,1)	16.489
Punjab	(2,1,1)	13.6	(2,1,1)	6.76	(3,1,1)	16.499
Balochistan	(1,1,1)	13.3	(1,1,1)	7.01	(1,1,1)	17.162
KPK	(2,1,1)	12.3	(2,1,1)	7.04	(2,1,1)	17.166
Islamabad	(1,1,1)	10.3	(1,1,1)	2.84	(1,1,1)	10.477
AJK	(2,1,1)	10.2	(3,1,1)	2.81	(6,1,1)	10.513
GB	(1,1,1)	11.4	(1,1,1)	5.19	(1,1,1)	13.574

Table 6: Coefficients of ARIMA (p, d, q) confirmed cases

Area	C	AR(p)	MA(q)
S ARIMA(1,1,1)	2.527896	-0.198979	-0.329277
P ARIMA(1,1,1)	1.420884	-0.084796	-0.535507
B ARIMA(1,1,1)	0.031200	0.122658	-0.718450
KPK ARIMA(1,1,1)	0.768481	-0.051152	-0.595540
I ARIMA(1,1,2)	0.230501	-0.002968	-0.484353
AJK ARIMA(1,1,1)	0.079609	0.033506	-0.741501
GB ARIMA(1,1,1)	0.002563	-0.049739	-0.775019

S: Sindh, P: Punjab, B: Balochistan and I: Islamabad

Table 7: Coefficients of ARIMA (p, d, q) death cases

Area	C	AR(p)	MA(q)
S ARIMA(1,1,1)	0.037185	0.046475	-0.787666
P ARIMA(1,1,1)	0.068544	-0.250547	-0.588371
B ARIMA(3,1,1)	0.000953	-0.199120	-0.781526
KPK ARIMA(2,1,1)	0.027085	0.056550	-0.774623
IARIMA(2,1,1)	0.003467	-0.059009	-0.816307
AJK ARIMA(2,1,1)	0.003808	-0.044328	-0.874293
GB ARIMA(1,1,1)	0.242180	0.986371	-0.933573

Table 8: Coefficients of ARIMA (p, d, q) recovery cases

Area	C	AR(p)	MA(q)
S ARIMA(1,1,1)	2.559267	0.228852	-0.885770
P ARIMA(1,1,1)	1.920634	0.077359	-0.934492
B ARIMA(1,1,1)	0.098233	0.535011	-0.873252
KPK ARIMA(1,1,1)	0.866882	0.279519	-0.891849
I ARIMA(1,1,2)	0.367730	-0.906239	-0.603769
AJK ARIMA(1,1,1)	0.056927	-0.119501	-0.823448
GB ARIMA(1,1,1)	0.006511	-0.079661	-0.897576

Table 9: Correlation of all provinces of Pakistan actual data and forecast of July–August 20201

Area	Jul- Aug 2021 CC	Jul- Aug 2021 DC	Jul- Aug 2021 RC
S	-0.7064	-0.6444	-0.1658
P	0.8236	0.5164	0.4458
B	0.5586	0.1600	0.3681
KPK	0.7309	0.5507	0.3465
I	0.8121	0.5511	0.7450
AJK	-0.4813	0.1755	0.8263
GB	0.1799	0.0982	-0.0856

4.3. Neighboring country comparison

Some of the neighboring countries are compared for confirmed and death cases like India, Bangladesh, and Iran in Figs. 5 and 6. It is observed that India has a higher rate of confirmed and death cases than Iran. Pakistan and Bangladesh have the almost same rate of COVID-19 cases for confirmed and death. In addition, it's to make mention of similar studies (Ghosh, 2020; Fargana et al., 2021). According to WHO in India, Bangladesh, and Iran DC cases are reported 2893589, 146020, and 843140 respectively. The graphs clearly show that there is a higher side of confirmed cases and death cases in India whereas; Bangladesh manages a uniform low rate in both of these cases. A better economy and better literacy rate perhaps have played their role in better control. In India dense population may be a possible cause of the high spread of the disease. Table 10 shows LRM variable results for total C, R, and D cases.

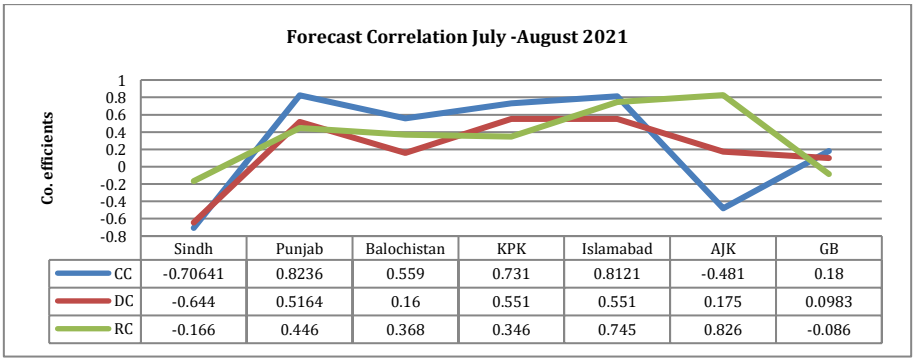


Fig. 4: Correlation of August 2021 with actual and forecasted data

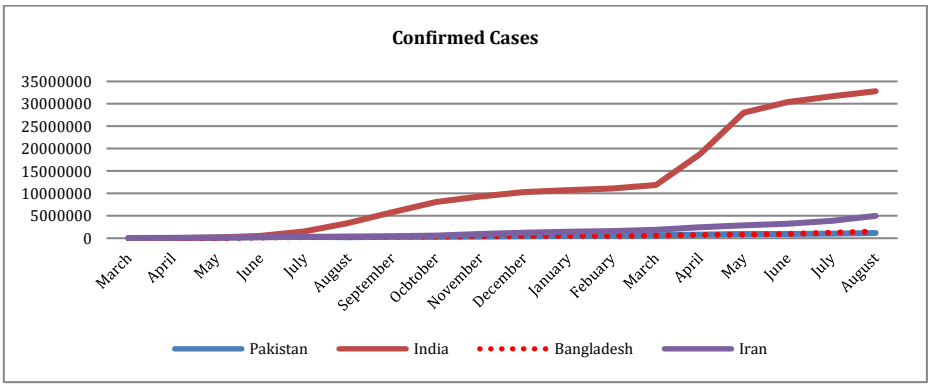


Fig. 5: Confirmed cases of neighboring countries (March 2020 to August 2021)

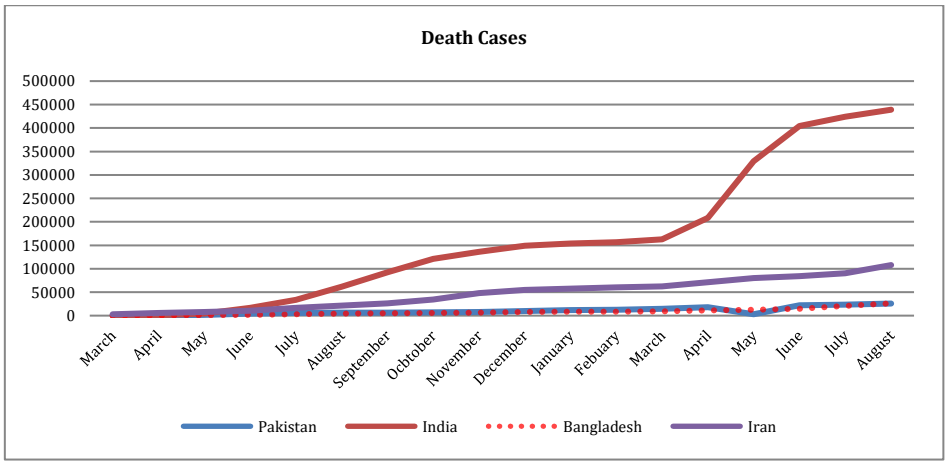


Fig. 6: Death cases of neighboring countries (March 2020 to August 2021)

Table 10: LRM variable results for total C, R, and D cases

Variables	(M)	(T)	(H)
Total Confirmed Cases			
a_o	12103.110	112480.586	-168929.355
b_o	7868.033	-2423.185	4000.712
R^2	0.241	0.097	0.129
Adj R^2	0.177	0.022	0.057
Standard E	0.252	0.274	0.269
Total Recovered Cases			
a_o	13202.176	86075.852	-220755.272
b_o	6420.996	-1447.811	4701.915
R^2	0.198	0.043	0.221
Adj R^2	0.132	-0.037	0.156
Standard E	0.258	0.282	0.255
Total Death Cases			
a_o	206.681	2469.610	-4303.383
b_o	173.862	-56.174	96.900
R^2	0.258	0.114	0.166
Adj R^2	0.196	0.041	0.097
Standard E	0.249	0.272	0.264

M: months, T: Temperature, and H: humidity

4.4. Regression analysis

The R^2, adjusted R^2, standard error, and coefficients of the model have been observed for the best fit model is presented in Table 10 for LRM.

Figs. 7 to 9 presents the linear regression for the total confirmed (Fig. 7), totally recovered (Fig. 8), and total death cases (Fig. 9) to show the predicted statistics with respect to the monthly data of increasing cases (Figs. 7-9 a and d), the effect of average temperature (Figs. 7-9 b and e), and effect of average humidity (Figs. 7-9 c and f), using 95% confidence interval. It has been observed the value of the R^2, adjusted R^2 for temperature and humidity is smaller than monthly growth. Affected cases increases as humidity increases and the temperature do not have a significance effect on the number of cases.

Fig. 10 presents the multi regression modeling with similar coefficients in Table 10. It has been observed that the numbers of confirmed and recovered cases are increasing on the monthly basis in Fig 10a. Fig. 10c shows the humidity analysis has been confirmed and the confirmed cases increase with the rise in humidity. Fig. 10b indicates that temperature does not directly affect the cases, as the regression line decreases when the temperature increases. Fig. 11 presents the cumulative exponential growth function as nonlinear regression analysis. The graphs are presented with trend equations on the basis of monthly growth in Fig. 11 (a, b, and c). EGM is performed similarly to linear regression for the temperature and humidity as shown in Figs. 7, 8, and 9 (d, e, and f). EGM shows the better performance of the LRM and MR (error increases); it shows the fewer data fluctuations across the regression line and has the smaller standard error.

5. Conclusion

The presented study has been performed on the exploratory data analysis to understand the behavior of confirmed cases, death cases, and recovery cases in Pakistan to investigate the spread during different waves. Moreover, the autoregressive

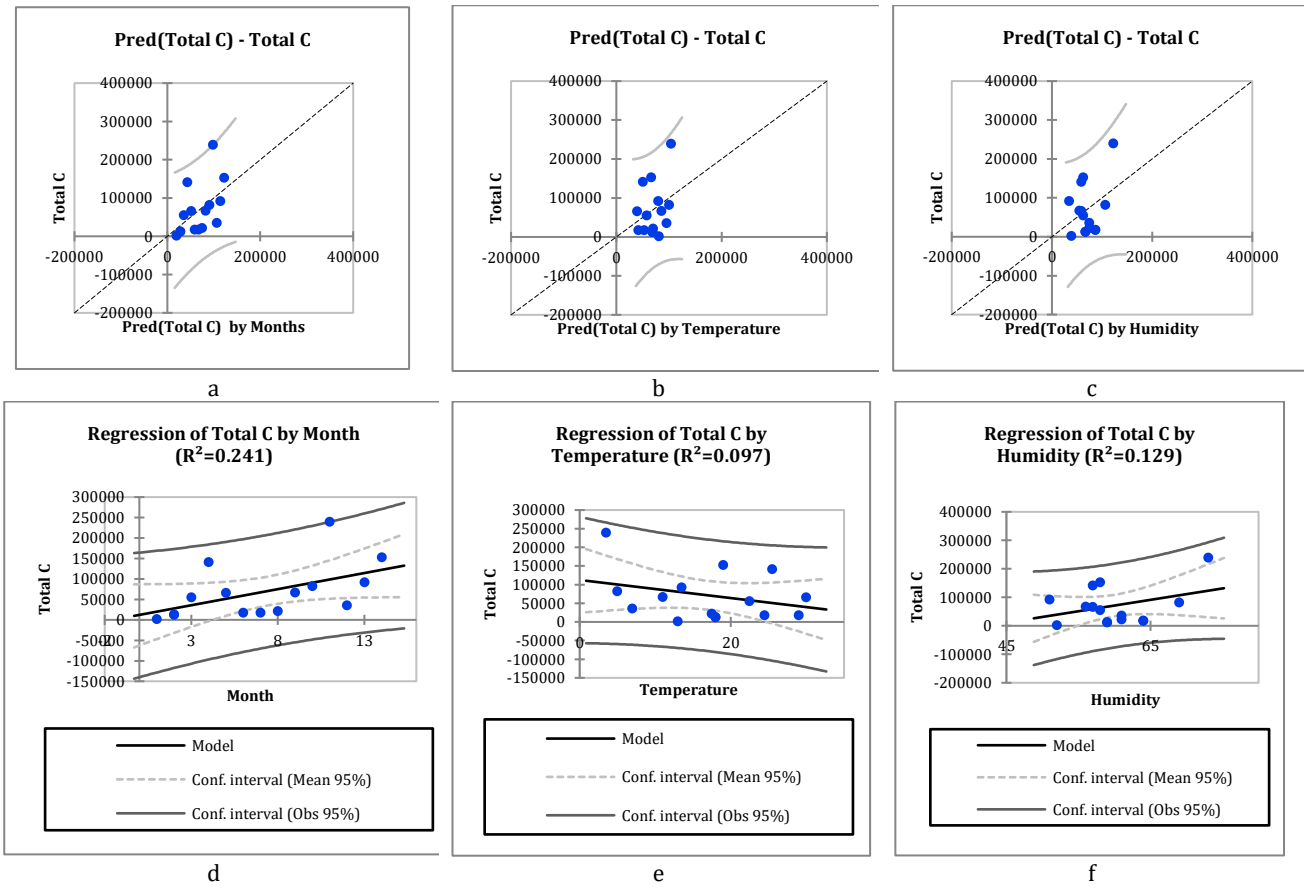


Fig.7: Predicted values of total confirmed cases and linear model with 95% confidence interval, prediction of confirmed cases with a) months, b) temperature, c) humidity, and (d, e, and f) respective linear regression models

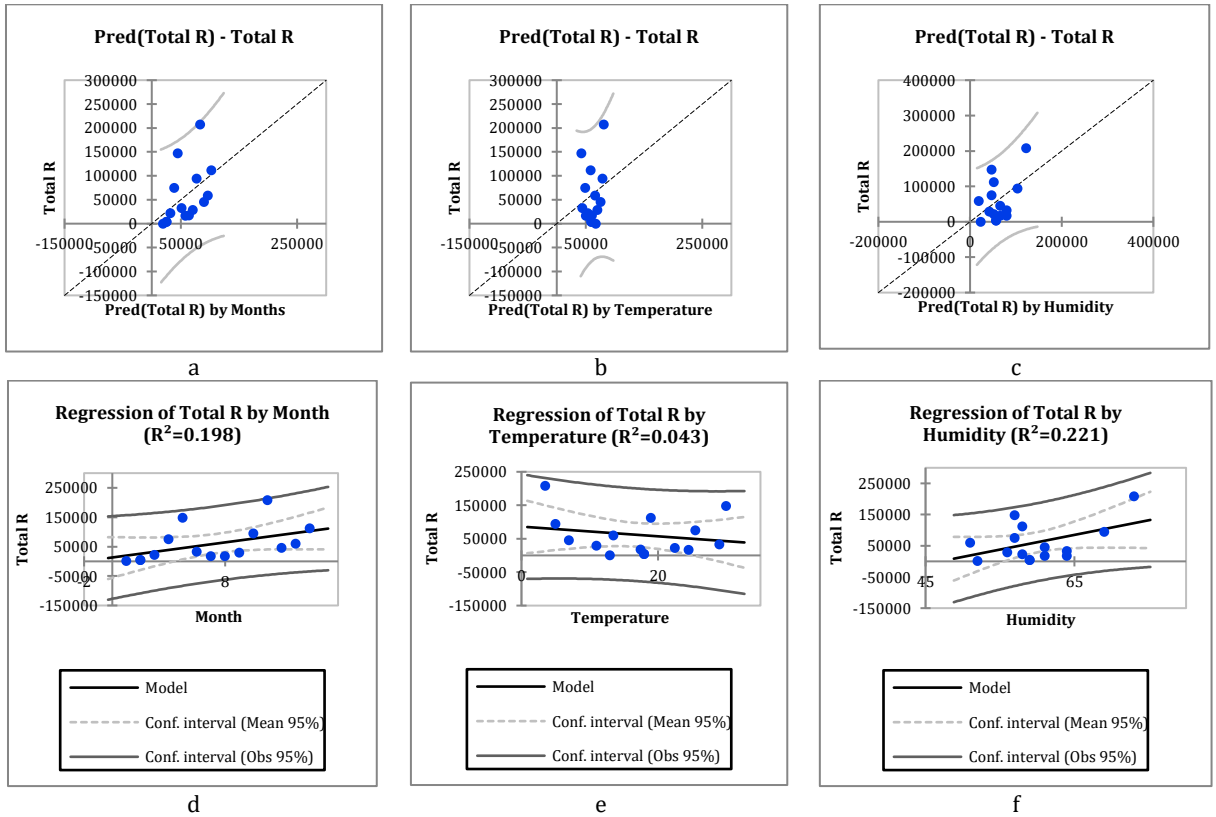


Fig. 8: Predicted values of total recovered cases and linear model with 95% confidence interval, prediction of recovered cases with a) months, b) temperature, c) humidity, and (d, e, and f) respective linear regression models

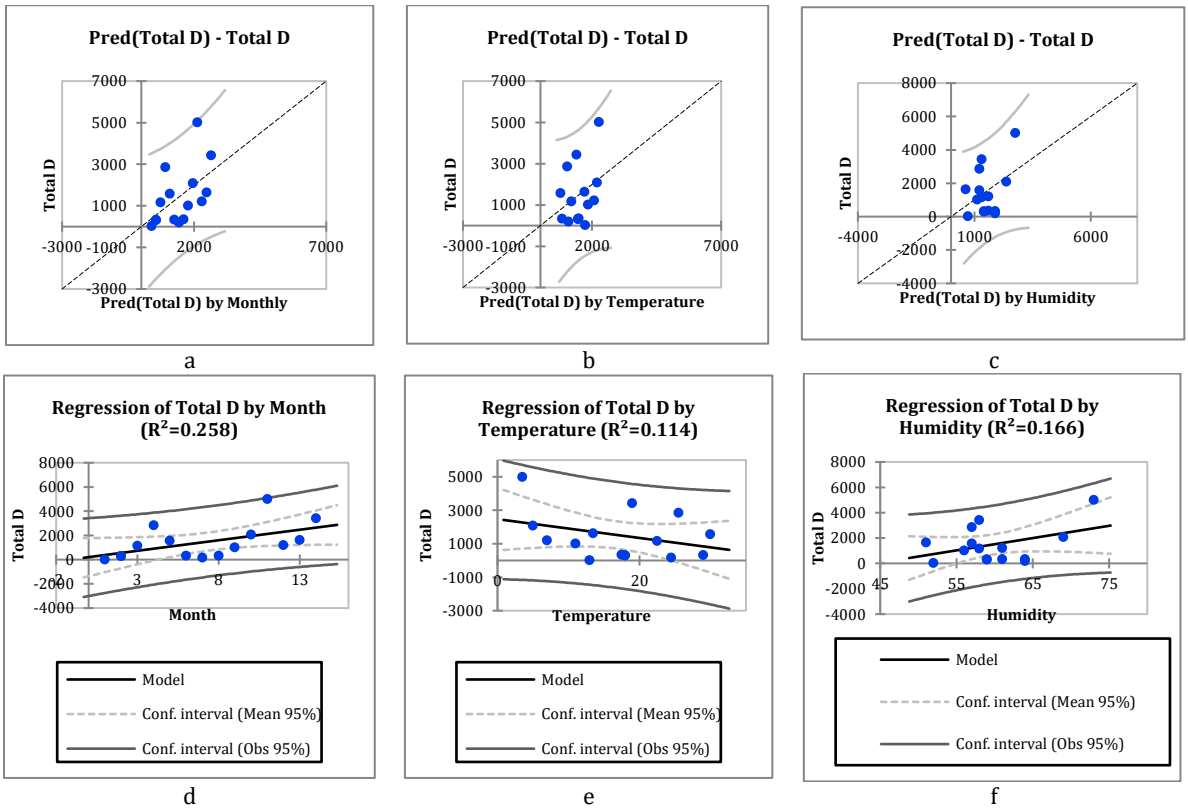


Fig.9: Predicted values of total death cases and linear model with 95% confidence interval, prediction of death cases with a) months, b) temperature, c) humidity, and (d, e, and f) respective linear regression models

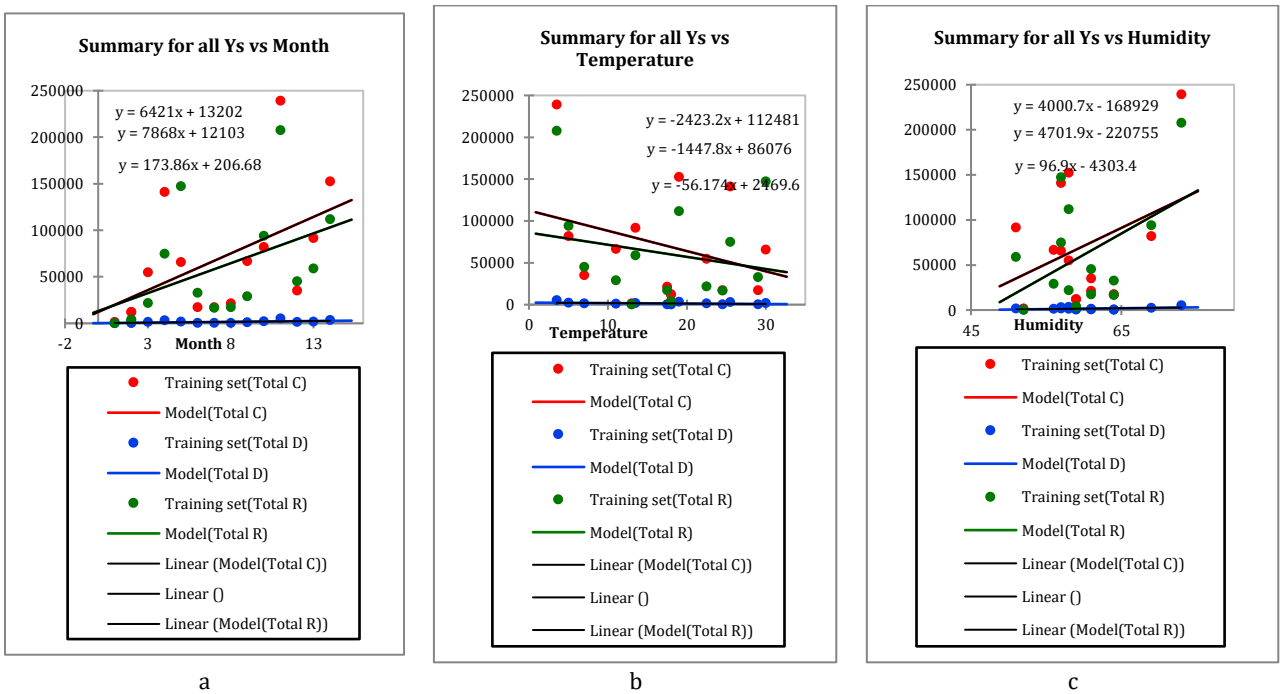
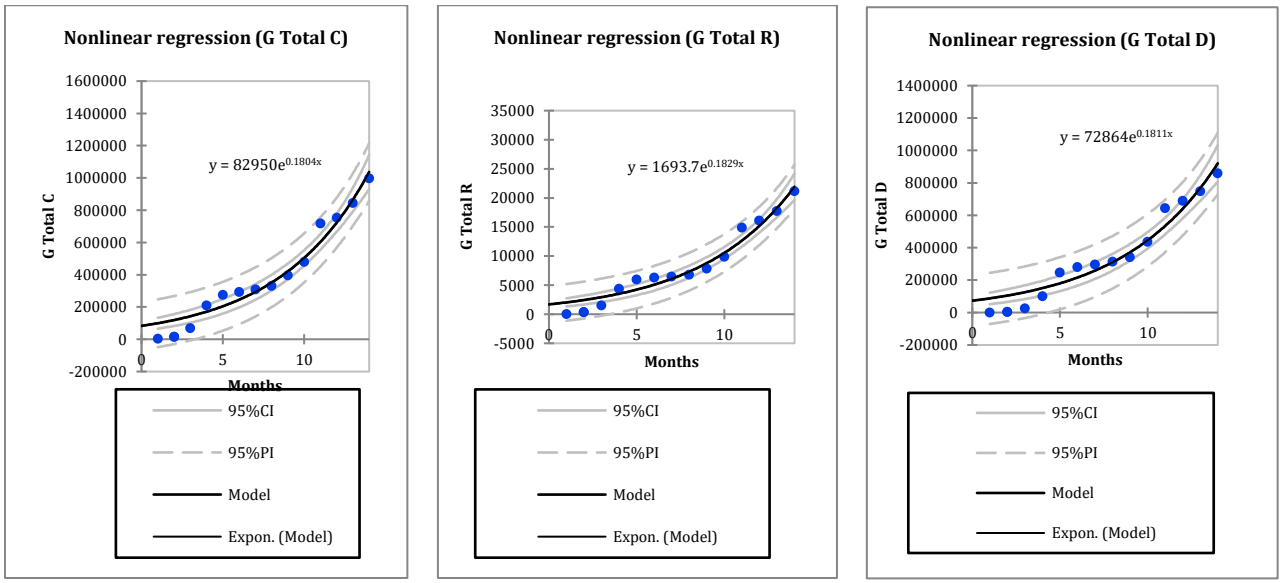


Fig.10: Multiple regressions for confirmed, recovered, and death with a) months, b) temperature, c) humidity

An integrated moving average model has been applied, which includes model identification followed by model estimation with diagnostic checking and future trend predicted for COVID-19 through correlation function. The best models for CC, RC, and DC have been identified. It has been observed that populated areas are more affected and the death rate increases as well. Also, the number of recoveries improved during the later waves. Moreover, exploratory data analysis has been performed on the basis of monthly increasing cases, average temperature, and humidity. The descriptive and graphical statistics have been analyzed with the models LRM, MR, and EGM. It has been observed that fluctuations across the regression line for LRM and EM models, it is detected that cases increases as the humidity increases and the EGM is found better adequate to predict the COVID-19 growth.



a b c

Fig. 11: Exponential growth function with respect to the number of months, a) Confirmed cases, b) Recovered cases and c) Death cases

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Anderson RM, Heesterbeek H, Klinkenberg D, and Hollingsworth TD (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? The Lancet, 395(10228): 931-934. https://doi.org/10.1016/S0140-6736(20)30567-5

Fargana A, Arifutzzaman A, and Rakhimov AA (2021). Spreading analysis of COVID-19 epidemic in Bangladesh by dynamical mathematical modelling. European Journal of Medical and Educational Technologies, 14(3): em2109. https://doi.org/10.30935/ejmets/10959

Ghosh S (2020). Predictive model with analysis of the initial spread of COVID-19 in India. International Journal of Medical Informatics, 143: 104262. https://doi.org/10.1016/j.ijmedinf.2020.104262 **PMid:32911257 PMCid:PMC7445130**

Gunnell D, Appleby L, and Arensman E et al. (2020). Suicide risk and prevention during the COVID19 pandemic. The Lancet Psychiatry, 7(6): 468-471. https://doi.org/10.1016/S2215-0366(20)30171-1

Guo J, Chen H, and Chen S (2020). Improved kernel recursive least squares algorithm based online prediction for nonstationary time series. IEEE Signal Processing Letters, 27: 1365-1369. https://doi.org/10.1109/LSP.2020.3011892

Hamzah FB, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, and Chung MH (2020). CoronaTracker: Worldwide COVID-19 outbreak data analysis and prediction. Bulletin of World Health Organisation. https://doi.org/10.2471/BLT.20.255695

Imai N, Cori A, Dorigatti I, Baguelin M, Donnelly CA, Riley S, Ferguson NM (2020). Report 3: Transmissibility of 2019-nCoV. Imperial College London COVID-19 Response Team, London, UK.

Jung SM, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, and Nishiura H (2020). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. Journal of Clinical Medicine, 9(2): 523. https://doi.org/10.3390/jcm9020523 **PMid:32075152 PMCid:PMC7074479**

Kane MJ, Price N, Scotch M, and Rabinowitz P (2014). Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics, 15(1): 1-9. https://doi.org/10.1186/1471-2105-15-276 **PMid:25123979 PMCid:PMC4152592**

Kawohl W and Nordt C (2020). COVID-19, unemployment, and suicide. The Lancet Psychiatry, 7(5): 389-390. https://doi.org/10.1016/S2215-0366(20)30141-3

Khajanchi S, Das DK, and Kar TK (2018). Dynamics of tuberculosis transmission with exogenous reinfections and endogenous reactivation. Physica A: Statistical Mechanics and its Applications, 497: 52-71. https://doi.org/10.1016/j.physa.2018.01.014

Li XL (2020). Convolutional PCA for multiple time series. IEEE Signal Processing Letters, 27: 1450-1454. https://doi.org/10.1109/LSP.2020.3016185

Read JM, Bridgen JRE, Cummings DAT, Ho A, and Jewell CP (2021). Novel coronavirus 2019-nCoV (COVID-19): Early estimation of epidemiological parameters and epidemic size estimates. Philosophical Transactions of the Royal Society B: Biological Sciences, 376: 20200265. https://doi.org/10.1098/rstb.2020.0265 **PMid:34053269 PMCid:PMC8165596**

Sharma RR, Kumar M, Maheshwari S, and Ray KP (2020). EVDHM-ARIMA-based time series forecasting model and its application for COVID-19 cases. IEEE Transactions on Instrumentation and Measurement, 70: 1-10. https://doi.org/10.1109/TIM.2020.3041833

Shastri S, Singh K, Kumar S, Kour P, and Mansotra V (2020). Time series forecasting of COVID-19 using deep learning models: India-USA comparative case study. Chaos, Solitons and Fractals, 140: 110227. https://doi.org/10.1016/j.chaos.2020.110227 **PMid:32843824 PMCid:PMC7440083**

Talkhi N, Fatemi NA, Ataei Z, and Nooghabi MJ (2021). Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods. Biomedical Signal Processing and Control, 66: 102494. https://doi.org/10.1016/j.bspc.2021.102494 **PMid:33594301 PMCid:PMC7874981**

Tomar A and Gupta N (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Science of the Total Environment, 728: 138762. https://doi.org/10.1016/j.scitotenv.2020.138762 **PMid:32334157 PMCid:PMC7169890**

Wang YW, Shen ZZ, and Jiang Y (2018). Comparison of ARIMA and GM (1, 1) models for prediction of hepatitis B in China. PLOS ONE, 13(9): e0201987. https://doi.org/10.1371/journal.pone.0201987 **PMid:30180159 PMCid:PMC6122800**

WHO (2020). Coronavirus disease 2019 (COVID-19): Situation report, 73. World Health Organization, Geneva, Switzerland.

WHO (2021). Coronavirus disease (COVID-19). World Health Organization, Geneva, Switzerland.

Zareie B, Roshani A, Mansournia MA, Rasouli MA, and Moradi G (2020). A model for COVID-19 prediction in Iran based on China parameters. Archives of Iranian Medicine, 23(4): 244-248. https://doi.org/10.34172/aim.2020.05 **PMid:32271597**

Zhan C, Tse CK, Lai Z, Hao T, and Su J (2020). Prediction of COVID-19 spreading profiles in South Korea, Italy and Iran by data-driven coding. PloS One, 15(7): e0234763. https://doi.org/10.1371/journal.pone.0234763 **PMid:32628673 PMCid:PMC7337285**

Zhong L, Mu L, Li J, Wang J, Yin Z, and Liu D (2020). Early prediction of the 2019 novel coronavirus outbreak in the mainland China based on simple mathematical model. IEEE Access, 8: 51761-51769. https://doi.org/10.1109/ACCESS.2020.2979599 **PMid:32391240 PMCid:PMC7176026**