International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2023; 8(5): 47-52

https://www.mathsjournal.com
Received: 20-05-2023
Accepted: 27-06-2023

Ajibode IA
Department of Mathematics and Statistics, Federal Polytechnic Ilaro, Ogun State, Nigeria

Sikiru OA
Department of Mathematics and Statistics, Federal Polytechnic Ilaro, Ogun State, Nigeria

Corresponding Author:
Ajibode IA
Department of Mathematics and Statistics, Federal Polytechnic Ilaro, Ogun State, Nigeria

COVID-19 case modeling in Nigeria: Time series analysis using ARIMA and ARIMAX

Ajibode IA and Sikiru OA

Abstract
This study involved Modeling COVID-19 confirmed cases in Nigeria by using ARIMA and ARIMAX. The data of 33 months was collected based on confirmed cases and temperature, from Nigeria Center for Disease and Control and world weather information respectfully which spanned between February 2020 to October 2022. The aim of this study was to fit parsimonious models into the COVID-19 confirmed case data in the present of temperature. Stationarity of the data was achieved through first differencing. After thorough analysis, ARIMAX (0, 1, 2) with exogenous variable (temperature)performed better than ARIMA (0, 1, 2) RMSE and MAE.

Keywords: RMSE, MAE, ARIMA, ARIMAX, COVID-19, pandemic

1. Introduction
The emergence of the COVID-19 pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan, China, in December 2019 (Wu *et al*., 2020; WHO, 2020; Zhou *et al*., 2020; Huang *et al*., 2020; Zhu *et al*., 2020; Li *et al*., 2020; Lu *et al*., 2020; Gao, 2018; Wang *et al*., 2020) [17, 16, 21, 6, 22. 9, 10, 5, 15] has led to a global health crisis. This viral infection encompasses a wide range of respiratory ailments, from mild cold-like symptoms to severe diseases like Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS), characterized by fever, cough, pains, and weariness. Elderly individuals and those with underlying medical conditions are particularly susceptible to severe illness. Transmission of the virus primarily occurs through respiratory droplets when an infected person coughs or sneezes. To prevent its spread, numerous measures, including the closure of public institutions and restrictions on gatherings, have been implemented worldwide. In Nigeria, the first confirmed case of COVID-19 was reported in February 2020 (WHO, 2020; NCDC, 2020) [16].

Several researchers in Nigeria, Morocco, France, Italy, Spain, the United States, and China have used various techniques to predict COVID-19 deaths and incidence patterns.

The COVID-19 pandemic has prompted extensive research efforts worldwide, aiming to model and understand the transmission dynamics of the disease. Several studies in the existing literature have employed different modeling techniques to investigate the confirmed cases and spread of COVID-19 (Zhang *et al*., 2021; Youssef and Elhadary, 2021; Gaffar *et al*., 2021; Fanelli and Piazza, 2021; Ali *et al*., 2021; Di Carlo and Cappelli, 2021; Manrique-Molina and López-Parra, 2021; Nasa *et al*., 2021; Zhang, 2020; Sameni and Moslehi, 2021) [20, 18, 4, 3, 1, 2, 11, 12, 19, 13]. These studies collectively demonstrate the significance of statistical tools in gaining insights into the transmission dynamics of COVID-19 and aiding informed decision-making for pandemic management.

Statistical models and machine learning tools have been utilized to analyze and predict the spread of COVID-19. These techniques include time series analysis, compartmental models, Bayesian models, supervised learning, clustering and anomaly detection, and network analysis. They provide significant insights into the transmission of COVID-19 and can aid in making educated decisions about pandemic management. The present research paper endeavors to unravel the influence of temperature as an explanatory variable on the number of confirmed COVID-19 cases in Nigeria.

While numerous studies have investigated various factors affecting the spread of the virus, the incorporation of temperature as a significant variable remains largely unexplored within the existing literature. Through the utilization of advanced autoregressive integrated moving average (ARIMA) and ARIMAX models, this study aims to fill this knowledge gap and provide novel insights into the relationship between temperature and the prevalence of COVID-19 in Nigeria.

2. Methodology

The study used the ARIMA and ARIMAX models to determine the optimum model for describing confirmed cases of pandemic in Nigeria with inclusion of an exogenous variable temperature.

The study relies on secondary data collection, gathering 33 observations from two primary sources: the Nigeria Centre for Disease Control and the Nigeria weather prediction and information. The data spans from February 2020 to October 2022, providing a comprehensive temporal perspective for analysis.

2.1 Data Analysis: The initial stage in the data analysis method was to determine stationarity of the data series. This was achieved by using the Augmented Dickey Fuller (ADF) test was utilised, with three regression techniques available: no constant, no trend, constant without trend, and constant with trend. H_0: $\varphi=0$ vs H_1: $\varphi\neq0$ was the hypothesis set, and the decision rule was to reject H_0 if the p-value was larger than the level of significance (0.05). The time series data was then modelled using Auto-Regressive Integrated Moving Average (ARIMA).

2.2 Augmented Dickey Fuller (ADF) test

The ADF is used to test the unit test, with the assumption that u_t is a white noise error that is not autocorrelated. When the p-value exceeds the level of significance (5%), the null hypothesis is rejected.

2.3 ARIMA Model

The model can be represented as ARIMA (p, d, q), where p is the order of the auto regressive process, d is the order of the data differencing for stationarity and q is the order of the moving average process.

Auto-Regressive (AR) model is specified as:

$$A_t = c + \mathbf{L}_1 A_{t-1} + \mathbf{L}_2 A_{t-2} + \cdots + \mathbf{L}_p A_{t-p} + k_t \quad (1)$$

A_t is the the series in the present period
L is the coefficients
A_{t-p} is the series at lag p and
k_t is the error term respectively.

Also, the moving average (MA) model is specified as:

$$A_t = c + \mathrm{M}_1 k_{t-1} + \mathrm{M}_2 k_{t-2} + \cdots + \mathrm{M}_p k_{t-p} + k_t \quad (2)$$

Where,
M is the coefficients
k_{t-q} is the error term at q lags respectively.

2.4 ARIMAX Model

ARIMAX model is made up of the ARIMA model and the exogenous variable written as

$$A_t = \beta x_i + \sum_{j=1}^{p} \boldsymbol{L}_j A_{i-j} + k_t + \sum_{j=1}^{q} M_j k_{i-j} \quad (3)$$

This can be re-express as:

$$\mathrm{L(L)}\, A_t = \beta x_i + \mathrm{M(L)}\, k_t \quad (4)$$

This can be written as:

$$A_t = \beta x_i + d_i \quad (5)$$

Where,

$$d_i = \sum_{j=1}^{p} L_j A_{i-j} + k_t + \sum_{j=1}^{q} M_j k_{i-j}$$

2.5 Ljung-Box for Model Independence

The Ljung-Box formula can be express as:

$$\mathrm{Q(m)} = \mathrm{n(n+2)} \sum_{i=1}^{k-1} \frac{r_i^2}{n-i} \quad (6)$$

Where:
Q(m) refers to Ljung-Box test statistic
n represents the number of residuals
r_i^2 represents the squared autocorrelation of the residuals at lag i
i is the lag at which the autocorrelation is calculated

If the calculated Q statistic exceeds the critical value, it suggests evidence of residual autocorrelation, indicating that the model's residuals are not independent. On the other hand, if the calculated Q (m) statistic is smaller than the critical value, it suggests that the residuals are independent, supporting the adequacy of the model in capturing the temporal dependence in the data.

The null hypothesis is expressed as:

H_0: $\rho_1 = \rho_2 = \ldots = \rho_m = 0$

2.6 Shapiro Wilk Test for Residual Normality

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (7)$$

Where:

W is the Shapiro-Wilk test statistic
x_i represents the i-th ordered sample value
a_i are the coefficients that are pre-computed based on n
$\bar{x}$ represents mean of the sample values

The Shapiro-Wilk test statistic W is compared to critical values from the Shapiro-Wilk distribution table or through approximation methods to determine whether the data significantly deviates from normality.

2.7 Mean Absolute Error (MAE)

The MAE offers a measure of how well the model captures variability in data and makes accurate predictions in the context of ARIMA modelling.

The Mean Absolute Error (MAE) is computed as:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| Y_i - \widehat{Y}_i \right| \quad (8)$$

Where
MAE stands for the mean absolute error
N stands for number of observations
Y_i is the actual value at time i
Y_i is the predicted value at time i

The MAE provides a straightforward and interpretable measure of forecasting accuracy. A lower MAE indicates the model is better at predicting the actual values of the time series.
By comparing the MAE of different ARIMA models or different parameter combinations within an ARIMA model, it is possible to identify the model or parameter set that produces the most accurate predictions.

2.8 Root Mean Squared Error (RMSE)

RMSE is another commonly used for evaluating the performance of a model, which is also applicable to ARIMA models. It is a measure the magnitude of the forecasting errors, similar to MAE.
The RMSE is calculated by taking the square root of the average of the squared differences between the predicted values and the actual values of a time series. The formula for calculating RMSE is as follows:
It can be expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \quad (9)$$

Where P_i is the predicted value and O_i is observed value.

3. Discussion of Findings

Table 1: Description statistics of COVID-19 confirmed cases and temperature

Description statistics	Confirmed Cases	Temperature
Minimum	1	23.60
1rd Quarter	1260	24.60
Median	4044	25.50
mean	8065	25.85
3rd Quarter	11510	27.29
Maximum	43981	28.80

***Source*:** R Studio output

The above presentation descriptive statistics on monthly verified COVID-19 cases and temperature readings. The data, which spans from February 2020 to October 2022, shows that the maximum number of confirmed cases reported in a single month was 43,981, with an average of 8,065 affected people.

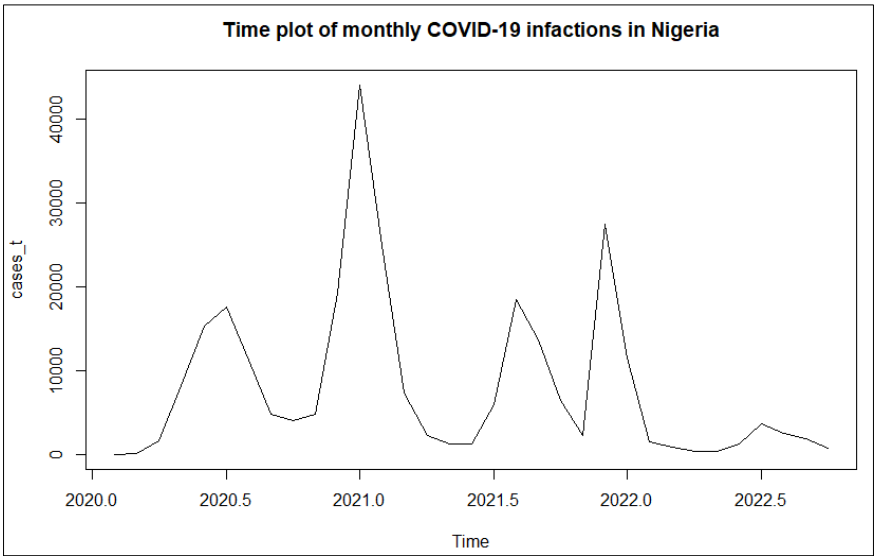

Fig 1: Monthly Covid-19 Confirmed

Figure 1 depicts the number of cases due to the virus. Non-stationarity of the data set is evidenced in the figure with no element of seasonality; this is exhibited in the plots of figures 2 and 3.

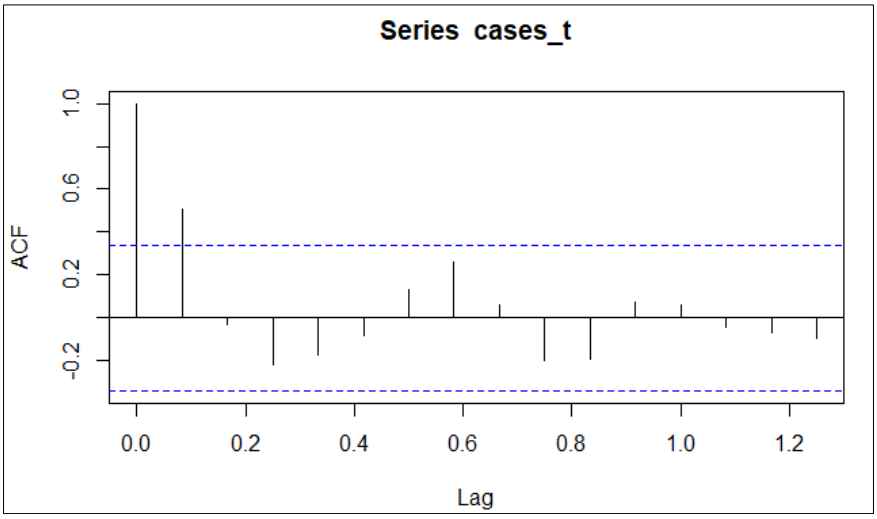

Fig 2: ACF plot of Monthly COVID-19 Confirmed

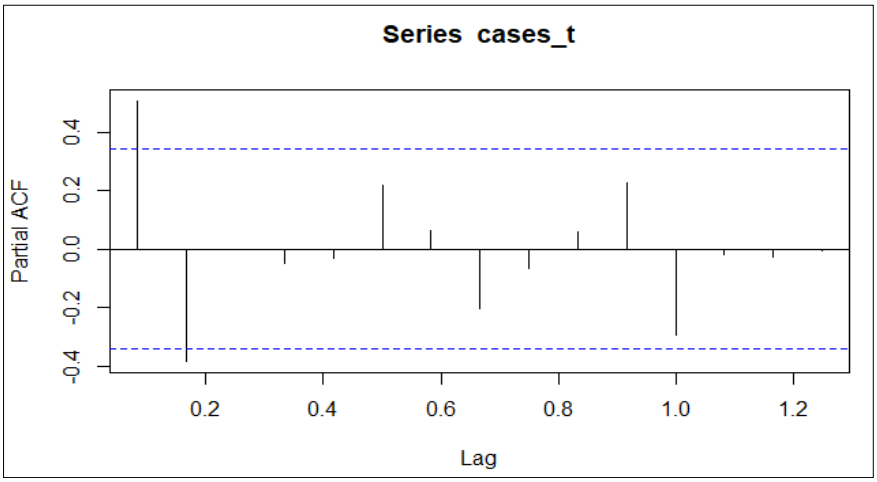

Fig 3: PACF plot of Monthly COVID-19 Confirmed

The spike of the ACF and PACF plots affirmed that confirm cases is non-random. The non-randomness could be attributed to different measures put in place by respective states in combating the pandemic.

Table 2: Unit Root Test

Variable	ADF @ Levels	P-valve	ADF @ First Difference	P-value	Remarks
Confirmed	-3.3409 3*	0.06879	-3.9774[3]**	0.02268	I(1)

ADF Critical Value at 5% = -2.95;
3* Indicates that a maximum lag length of 3 was included in the tests.
** indicates significant at 5%
***Source*:** Extracted from R-Studio Output

Presence of unit root is evidenced in table 2, coupled with the Dickey-Fuller test and p-values exceeding the 0.05 significance level. This result suggest that the data exhibited unit root, hence, differencing is necessary for model development. First-order differencing was applied to the series with an evidence of stationarity and confirming it as an order of I(1). The ACF and PACF plots below (figure 4 and 5) show an indication of the stationarity of the data.

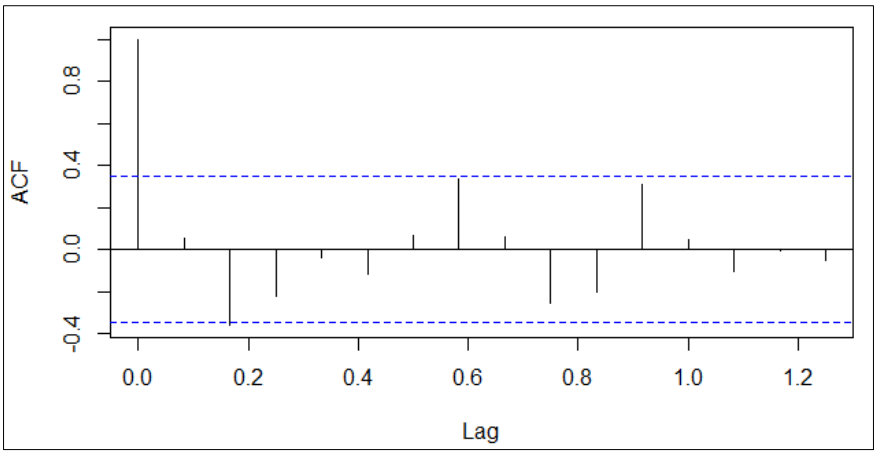

Fig 4: ACF plot of First Differenced Monthly COVID-19 Confirmed Cases

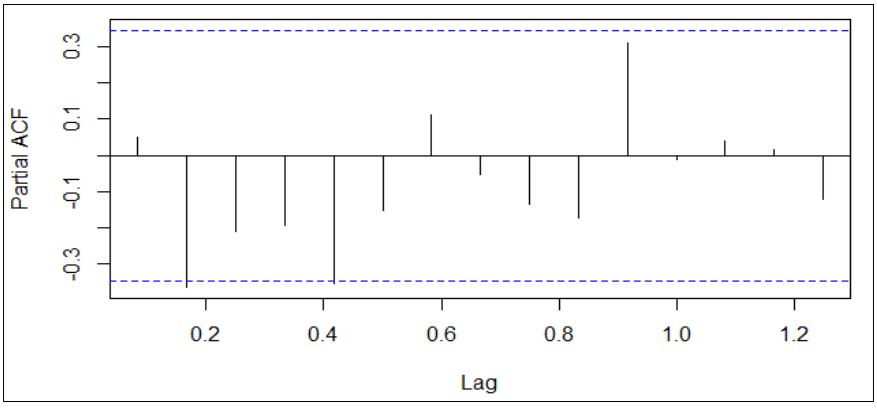

Fig 5: PACF plot of First Differenced Monthly COVID-19 Confirmed Cases

First differencing result indicates stationarity of the data set with significant spikes at the respective lags.
Because of the data set's stationarity, alternative ordering for the model were explored, and the results are shown in table 3 below.

Table 3: Iterated AIUMA Models for selection using Nigeria COVID-19 Confidned Cases

SIN	Modals	Variance	Log Likelihood	AIC	BIC
1	ARIMA(0,0,1)	67501950	-343.38	692.96	697.480
2	ARIMA(2,1,2)	64483840	-333.79	677.57	684.903
3	ARIMA(0,1,2)	68211088	-334.82	675.65	680.044
4	ARIMA(1,1,2)	63620437	-334.24	676.48	682.344
5	AR1MA(1,1,1)	90256190	-338.52	683.04	687.434
6	ARIMA(3,1,1)	64269884	-333.80	677.59	684.920
7	ARIMA(2,1,0)	80379911	-336.77	679.54	683.938
8	ARIMA(0,1,1)	92264609	-338.86	681.71	684.646
9	ARIMA(1,1,0)	92738252	-338.93	681.86	684.795

The presented table contains multiple ARIMA models of varying orders, with the goal of selecting the best model based on the lowest AIC and BIC values. Following from the result presented in table 3, the model with an order of (0, 1, 2) provides the better fit for the confirmed COVID-19 pandemic cases in Nigeria.
The table below illustrates the ARIMA orders and their associated parameter assessments for significance, following the principle of parsimony, which states that the simplest model should be chosen as long as it meets all basic requirements.

Table 4: ARIMA (0,1,2) Model Estimates

Parameters	Estimates	Standard Error	*Z value*	Pr(>|Z|)
L_1 MA(1)	-03260	01846	-17660	0.07739#
$L2$ MA(2)	-0.5684	0.1722	-3.3011	0.00096**

**significance of the coefficients @1% level
* significance of the coefficients @10% level
RMSE = 8132.901, Source: R-Studio Output

The model can be specified as:

$$\hat{X}_t = X_{t-1} - X_{t-2} - 0.3260\varepsilon_{t-1} - 0.5684\varepsilon_{t-2} \quad (10)$$

The estimated parameters of AR (0), MA (2) are as presented in the table and they statistically significant at 10% and 1% level of significance respectively. This is an evidence that the confirmed cases of COVID-19 pandemic can well be describe using ARIMA (0, 1, 2) model.

Table 5: ARIMAX Model Estimates

Parameters	Estimates	Standard Error	*Z value*	Pr(>|Z|)
L_1 MA(1)	-0.34261	0.15965	-2.1460	0.031875#
L_2 MA(2)	-0.47442	0.15268	-3.1072	0.001889##
Temperature	-2819.1207	1120.6223	-2.5157	0.011880#

Significance of the coefficients @I% level
Significance of the coefficients @5% level
RMSE = 7528.575, Source: R-Studio Output

The ARIMAX model fitted can as well be written as:

$$X_t = -0.34261\varepsilon_{t-1} - 0.47442\varepsilon_{t-2} - 2819.1207(temperture) \quad (11)$$

Where X_t is the number of confirmed cases due to COVID-19 pandemic on monthly basis.

The inclusion of temperature as an exogenous variable stemmed from the recognition that temperature could potentially influence the reported number of cases in Nigeria (Smith, 2020) [14]. Researchers observed that individuals residing in regions characterized by higher temperatures, which are more susceptible to infection, might have a lower likelihood of contracting the virus (Jones *et al*., 2019) [8]. As a result of this, temperature readings was fitted into the ARIMA (0,1,2) as an exogenous variable. The selected order of the AR and MA were found to significantly contribute to the ARIMAX model (p-value < 0.5), including the exogenous variable (p-value < 0.05), hence contributing to the model as it indicated that a unit increase in temperature reduces the number of confirmed cases in Nigeria.

Table 6: ARIMI and AR1M4X modals Shapiro-Wilk Test of Residual Normality

Models	Shapiro-Wilk	p-value
ARIMA	0.81917	0.8156
ARIMAX	0.93567	0.0828

***Source*:** R-Studio Output

The results of the Shapiro-Wilk normality test (table 6) show that the residuals of the ARIMA and ARIMAX models are normally distributed. ARIMA and ARIMAX have test statistic values of 0.81917 and 0.93567, respectively, and both have corresponding p-values larger than the 0.01 level of significance. As a result, at the 1% significance level, the hypothesis that the residuals are not normally distributed can be rejected.

Table 7: Ljung-Box Portmanteau Test

Models	Chi-Squared	Degree of freedom	p-value
ARIMA	0.40515	1	0.5244
ARIMAX	0.21317	1	0.6443

***Source*:** R-Studio Output

The Chi-square and Box tests performed on the residuals of the fitted ARIMA and ARIMAX models yielded significant results. The test statistic had a high value, yet the accompanying p-values were substantial, preventing the null hypothesis from being rejected, implying that the autocorrelation functions are zero. As a result, the residuals showed no significant non-zero autocorrelations, showing that the models successfully captured the series' dependence.
Additionally, the AIC and Log-likelihood values, which assess the model's goodness of fit and parsimony, revealed its efficacy and capacity to create correct predictions. As a result, both the ARIMA and ARIMAX models are thought to be suitable for forecasting the pattern of confirmed coronavirus infections in Nigeria, particularly when temperature is included as an exogenous variable.

Table 8: Predictions Evaluation for COVID-19 Confirmed Cases

Models	MAE	RMSE
ARIMA	5617.272	8132.901
ARIMAX	5303.113	7528.575

***Source*:** R-Studio Output

The ARIMAX model was shown to be the best prediction model for COVID-19 in the study, as evidenced by its superior performance in terms of root mean squared error (RMSE) and mean absolute error (MAE). The addition of temperature as an exogenous variable improved the ARIMAX model's

performance even more, validating its selection as the best choice for forecasting the virus's spread in the study.

4. Conclusion

According to the study's findings, an analysis of monthly confirmed COVID-19 cases from February 2020 to October 2022 found a peak average of 43,981 confirmed cases in a given month, with an average of 8,065 infected persons. Furthermore, the study revealed that temperature had a significant impact on the likelihood of developing the condition. This discovery is consistent with the findings of Ibrahim *et al*. (2021) [7], indicating the importance of temperature in COVID-19 transmission.

Furthermore, when model selection measures such as AIC, BIC, MAE, and RMSE were compared, the ARIMAX model beat the ARIMA model.

5. References

1. Ali N, Kerdprasop N. Analysis and forecasting of COVID-19 spreading in Asian countries. Chaos, Solitons & Fractals. 2021;139:110329.
2. Di Carlo P, Cappelli M. Using the SEIR model to analyze the spatiotemporal transmission of COVID-19 in Italy. Chaos, Solitons & Fractals. 2021;142:110418.
3. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals. 2021;139:110060.
4. Gaffar A, Yasmin M, Rahman A, Hossain MB. Modeling and forecasting of COVID-19 spread in Bangladesh: A comparison of statistical models. Chaos, Solitons & Fractals. 2021;142:110440.
5. Gao GF. From “A”IV to “Z”IKV: Attacks from emerging and re-emerging pathogens. Cell. 2018;172(6):1157-1159.
6. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395(10223):497-506.
7. Ibrahim NK, Abdelrahman AH, Salih AM, Eldein ES. Temperature and COVID-19: A study of two climatic zones in Sudan. Infection Ecology & Epidemiology. 2021;11(1):1924633.
8. Jones NR, Qureshi ZU, Temple RJ, Larwood JP, Greenhalgh T, Bourouiba L. Two metres or one: What is the evidence for physical distancing in COVID-19? BMJ. 2019;370:3223.
9. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, *et al*. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. New England Journal of Medicine. 2020;382(13):1199-1207.
10. Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. Journal of Medical Virology. 2020;92(4):401-402.
11. Manrique-Molina FJ, López-Parra JM. Forecasting COVID-19 epidemic in Colombia: A case study using a modified version of the classical SEIR model. Chaos, Solitons & Fractals. 2021;142:110428.
12. Nasa A, Thulasiram RK, Jeon G. Time series prediction of COVID-19 by mutation rate analysis using recurrent neural networks. Computers in Biology and Medicine. 2021;132:104338.
13. Sameni R, Moslehi G. Prediction of COVID-19 cases for top affected countries: An integrated approach using epidemiologic models and machine learning. PLoS ONE. 2021;16(9):e0257212.
14. Smith SM. Temperature and the transmission of COVID-19. Environmental Research. 2020;193:110255.
15. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. The lancet. 2020 Feb 15;395(10223):470-3.
16. World Health Organization. Coronavirus disease (COVID-19) pandemic; c2020. https://www.who.int/emergencies/disease/novel-coronavirus-2019
17. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. Jama. 2020 Apr 7;323(13):1239-42.
18. Youssef SA, Elhadary YA. COVID-19 prediction using hybrid models: A comparative study. Neural Computing and Applications. 2021;33(8):3793-3803.
19. Zhang Y. Dynamics of COVID-19 epidemics: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) waves, guidelines, policies and the caseload in the USA. Chaos, Solitons & Fractals. 2020;140:110290.
20. Zhang Z. Modeling and forecasting the COVID-19 pandemic in the United States: A data-driven approach. Chaos, Solitons & Fractals. 2021;140:110225.
21. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 Mar;579(7798):270-3.
22. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, *et al*. A novel coronavirus from patients with pneumonia in China, New England Journal of Medicine. 2019;382(8):727-733.