International Journal of Research in Medical Sciences
Bahuguna A et al. Int J Res Med Sci. 2023 May;11(5):1727-1734
www.msjonline.org
pISSN 2320-6071 | eISSN 2320-6012
DOI: https://dx.doi.org/10.18203/2320-6012.ijrms20231344

Original Research Article

Comparison of exponential smoothing and ARIMA time series models for forecasting COVID-19 cases: a secondary data analysis

Abhinav Bahuguna[1], Akanksha Uniyal[1]*, Neha Sharma[2], Jayanti Semwal[2]

[1]Department of Biostatistics, [2]Department of Community Medicine, Himalayan Institute of Medical Sciences, Swami Rama Himalayan University, Jolly Grant, Dehradun, Uttarakhand, India

Received: 17 March 2023
Revised: 16 April 2023
Accepted: 17 April 2023

***Correspondence:**
Akanksha Uniyal,
E-mail: akankshabiostats@gmail.com

ABSTRACT

Background: In order to manage outbreaks and plan resources, health systems must be capable of accurately projecting COVID-19 case patterns. Health systems can effectively predict future illness patterns by using mathematical and statistical modelling of infectious diseases. Different methods have been used with comparatively good accuracy for various prediction goals in medical sciences. Some illustrations are provided by statistical techniques intended to forecast epidemic cases. In order to increase healthcare systems readiness, this study aimed to identify the most accurate models for COVID-19 with a high global prevalence of positive cases.
Methods: Exponential smoothing model and ARIMA were employed on time series datasets to forecast confirmed cases in upcoming months and hence the effectiveness of these predictive models were compared on the basis of performance measures.
Results: It was seen that the ARIMA (0,0,2) model is best fitted with smaller values of performance measures (RMSE=4.46 and MAE=2.86) while employed on the recent dataset for short duration. Holt-Winters Exponential smoothing model was found to be more accurate to deal with a longer period of time series based data.
Conclusions: The study revealed that working with recent dataset is more accurate to forecast the number of confirmed cases as compared to the data collected for longer period. The early-stage warnings through these predictive models would be beneficial for governments and health professionals to be prepared with the strategies at different levels for public health prevention.

Keywords: ARIMA, COVID-19, Exponential smoothing, Forecast, Predictive models, Time series analysis

INTRODUCTION

The severe acute respiratory syndrome (SARS-CoV-2) outbreak brought on by the novel coronavirus (COVID-19) has resulted in a "global pandemic" due to its unparalleled rate of global transmission. More than ten million people from 200 countries have been infected with SARS-CoV-2 infection.[1] In order to manage outbreaks and plan resources, health systems must be capable of accurately projecting COVID-19 case patterns. Health systems can effectively predict future illness patterns by using mathematical and statistical modelling of infectious diseases.[2] Different methods have been used with comparatively good accuracy for various prediction goals. Some illustrations are provided by statistical techniques intended to forecast epidemic cases. Multivariate linear regression, simulation models, time series, back propagation neural networks, and grey forecasting are a few examples.[3-10] Any evolution in epidemiology is determined and influenced by a variety of variables, specifically by a tendency toward randomness. In hindsight, the aforementioned statistical

methods are hard to generalize. Because of its simple design and practical implementation, the ARIMA model has been effectively used on a much bigger scale in a variety of sectors.[11] ARIMA models are selected because they can be used to study the short-term impacts of acute infectious infections, are a versatile class of models that can be used to fit a variety of trajectories, and have a strong body of literature supporting them.[12] Several researchers in South Korea, Thailand, Iran, China, Brazil, and Italy employed ARIMA models to predict the COVID-19 outbreak trends.[13,14]

Forecasting time-series data with seasonal patterns, trend or both at once is done using the smoothing method. When evaluating the worth of a specific year, smoothing is the process of calculating the average value over a number of years.[15] Both the exponential smoothing method and the smoothing technique are subcategories of the smoothing method. By assigning various weights for historical data that have exponential decreasing characteristics, the exponential smoothing approach can be used to forecast data that is affected by seasonal or trend patterns.[16]

According to a study, during the course of the following 30 days, accumulated confirmed cases and daily new cases in the Brazil, USA, and India will be predicted by the ARIMA, SARIMA, and Prophet models. They claimed that SARIMA is more likely to show over-fitting in the USA, whereas the Prophet model has a better advantage in the prediction of COVID-19 there. When it comes to the forecasting of new cumulative instances for Brazil and India, the ARIMA model is better able to fit and anticipate data showing a rising trend in diverse countries.[17]

A study conducted in Saudi Arabia aimed to accurately analyse and forecast the emergence of new cases of COVID-19 in order to create a framework for global pandemic preparedness and to slow the transmission of the disease. The objective of this study was to investigate daily new infections, recoveries, and fatalities using the Prophet Facebook machine learning technique and ARIMA statistical technique. Based on metrics for forecasting performance, it was discovered that both models performed well in predicting the time series of COVID-19 in Saudi Arabia with a slight edge going to the ARIMA model for forecasting capability and the Prophet model for a few hyper-parameters and simplicity.[18]

METHODS

The methodology adopted in this paper is described with the help of (Figure 1). This research study was based on COVID-19 confirmed cases extracted from health bulletin for all over the state of Uttarakhand from January 2021 to December 2022. The secondary data is publicly available at the official website (https://health.uk.gov.in) of Medical Health and Family Welfare (MoHFW), Government of Uttarakhand.[19]

Statistical analysis

The analysis was performed using R software version 4.2.2 after cleaning the whole dataset. Kwiatkowski-Phillips-Schmidt-Shin test (KPSS test) and Augmented Dickey Fuller test (ADF Test) were performed to check whether a given time series is stationary or not. Exponential smoothing and Auto regressive model (ARIMA) were employed on time series datasets to forecast the confirmed cases and the effectiveness of these predictive models were compared on the basis of performance measures.

Exponential smoothing

Exponential smoothing consists of three types; first type is single exponential smoothing, which is independent on trend and seasonality, and consider a single parameter α to generate fitted value and to forecast further.[12] It anticipates the most recent forecast. The mathematical equation for single exponential smoothing is given by:

$$St = \alpha Yt + (1- \alpha) St-1$$

Where, St is smoothed statistics, Yt is the actual record in time t, and α is smoothing constant ranges from 0 to 1.

The second type is Double exponential smoothing (also known as Brown's method) is best to deal with the data that possess trend with the help of two smoothed constants α and β, and produces multiple ahead forecasts. With trend and seasonality of a series taken into considerations, Triple exponential smoothing (also known as Holt Winter's method) is used which is based on three parameters α, β, and γ.

Auto regressive model (ARIMA)

Box and Jenkins (1970) or ARIMA is a forecasting technique which is composed by two terms: AR and MA, used for auto regression and moving average respectively.

The purpose of using ARIMA is to take trends as well random fluctuations into consideration. An ARIMA model is made up of three terms: p, d, and q. The expression for ARIMA model is as follows:

$$y_t=\theta_0+\phi_1 y_{(t-1)}\ldots\ldots+\phi_p y_{(t-p)}-\theta_1 €_{(t-1)}\ldots\ldots- \theta_q €_{(t-q)}$$

The order of AR term is represented by p, MR term by q, and d is the number of steps used in differencing to make series stationary. The autocorrelation (ACF) and partial autocorrelation (PACF) are used to estimate these parameters.

Model selection

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), were used to measure the accuracy of these time series models with the help of following formulas:

$$\text{RMSE} = \sqrt{\frac{(y_{i-}\hat{y}_i)^2}{n}}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_{i-}\hat{y}_i|$$

Where yi and $\hat{y}_i$, denote the observed and predicted values respectively. The lower values of these measures will yield better accuracy.

RESULTS

The process of time series analysis begin with visualizing the pattern of data. (Figure 2) shows the trend of confirmed cases since January 2021. Before applying time series models, the foremost step is to check the stationary condition of datasets. The ADF test and KPSS test were used for this purpose and the hypothesis is set up as follow.

Augmented Dickey-Fuller (ADF) test

Null hypothesis (H^0): the data for the confirmed cases is not stationary.

Alternative hypothesis (H^1): the data for the confirmed cases is stationary.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

Null hypothesis (H^0): the data for the confirmed cases is stationary.

Alternative hypothesis (H^1): the data for the confirmed cases is not stationary.

On applying Augmented Dickey-Fuller (ADF) test on 'dataset 1' (Table 1), the p-value is 0.028 (<0.05), implying that null hypothesis (H^0) is rejected at a 5% level of significance, showing that the data for the confirmed cases is stationary, whereas for Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, p-value is greater than 0.10 which mean the data is stationary at 10% level of significance. KPSS test statistic of confirmed cases for the last 90 days ('dataset 2') shows the data to be non-stationary (p-value <0.10) and after 1st order differencing the series became stationary (p-value>0.10) as represented in (Table 2). The non-stationary and stationary series are shown with the help of graphical representations in (Figure 3) and (Figure 4) respectively.

Table 1: Test for stationary: monthly time series data for last 2 years (dataset 1).

Statistic	Value	P-value
ADF test	-3.8989	0.02865
KPSS test	0.22901	> 0.10

Table 2: Test for stationary: recent data recorded for last 90 days (dataset 2).

Statistic	Before differencing		After differencing	
	Value	P-value	Value	P-value
ADF test	-4.2713	0.01	-6.86	0.01
KPSS test	1.0054	0.01 (<0.10)	0.0267	> 0.1

Table 3: Smoothing parameters for different time series models.

Exponential smoothing model	Monthly data (dataset 1)	Recent data (dataset 2)
Single exponential smoothing	α= 0.0185	α= 0.0094
Holt's linear trend model	α= 0.1693, β= 0.077	α= 0.0726, β= 0.4615
Holt-winters smoothing	α= 1, β= 0, γ= 0.1	-------

Forecasting model

Exponential Smoothing

Exponential smoothing is the most common forecasting time series model where recent observations are assigned more weights compared to that of past observations. The smoothing model is much preferred for short term forecasting. (Table 3) shows parameters of different exponential smoothing models when employed on two types of series.

Dataset 1: monthly based time series data recorded for last two years.

Dataset 2: daily recorded data on confirmed cases for last 90 days.

Single exponential smoothing is well suited when there is no trend or pattern in a series by taking one smoothing parameter (α) into considerations, Holt's linear trend (also called as, double exponential smoothing) is good to deal with the data which have linear trend by taking two smoothing parameters (α, β).

Holt-Winter model (also known as, triple exponential smoothing), with three smoothing parameters (α, β, γ), is best to deal with the series which have trend as well as seasonality.

Table 4: Forecasts results with 95% prediction interval for upcoming months (dataset 1).

Month	Single exponential smoothing		Arima model	
	Forecast	95 % prediction interval	Forecast	95 % prediction interval
Jan-23	8325	(-65929 - 82581)	26664	(-34428 - 87757)
Feb-23	8325	(-65942 - 82593)	4244	(-6108 - 14597)
Mar-23	8325	(-65955 - 82606)	2954	(-4694 - 10602)
Apr-23	8325	(-65968 - 82619)	57486	(-100107 - 215079)
May-23	8325	(-65981 - 82632)	104226	(-197665 - 406117)
Jun-23	8325	(-65993 - 82645)	6166	(-12671 - 25004)

Table 5: Parameters of ARIMA model (dataset 1).

Model	Log likelihood	AIC	BIC
ARIMA (0,0,1)	-285.14	574.27	576.63
ARIMA (0,0,2)	-260.12	526.24	533.71

Table 6: Evaluation of models based on performance measures.

Model	Monthly data (dataset 1)		Recent data (dataset 2)	
	RMSE	MAE	RMSE	MAE
Single exponential smoothing	37766.34	18757.61	5.42	3.68
Holt's linear trend model	40815.69	20834.52	5.91	3.98
Holt-Winters smoothing	9864.69	6584.48	-----	------
ARIMA (0,0,1)	34790.13	17155.16	-----	-------
ARIMA (0,0,2)	-------	-------	4.46	2.86

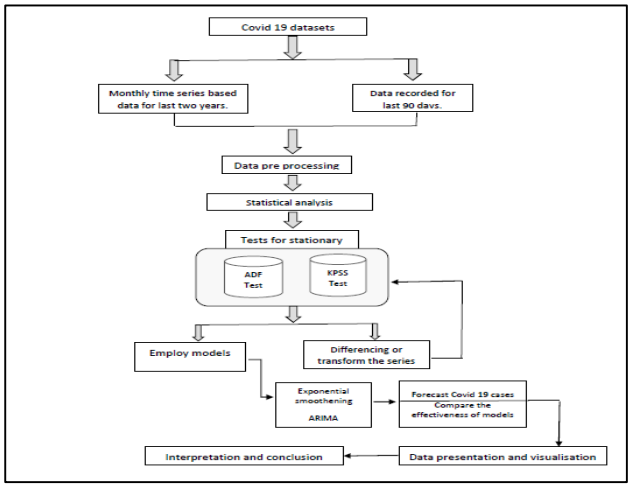

Figure 1: Flow chart for proposed methodology.

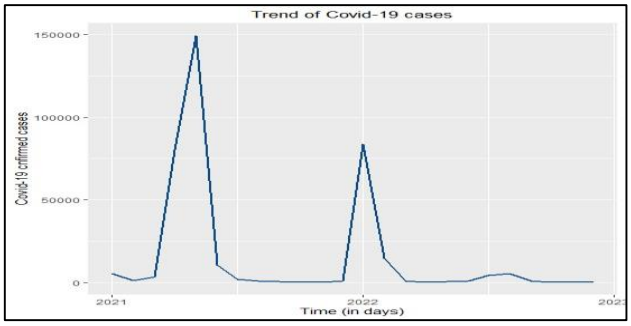

Figure 2: Trend of confirmed cases (Jan 2021 - Dec 2022): dataset 1.

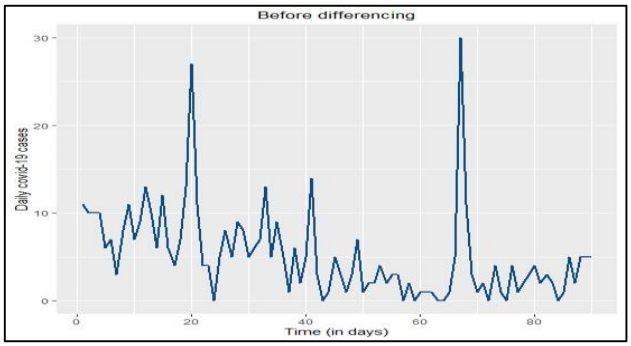

Figure 3: Trend of confirmed cases for last 90 days-before differencing: dataset 2.

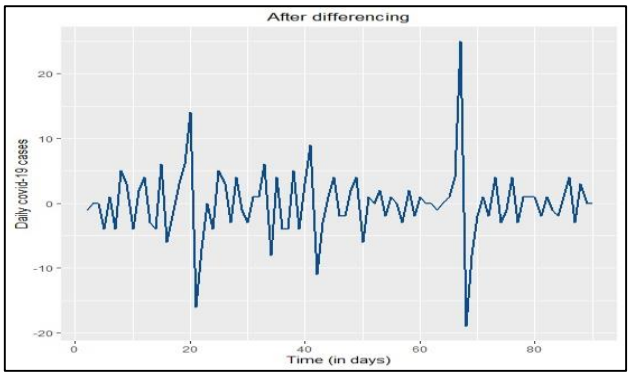

Figure 4: Trend of confirmed cases for last 90 days-after differencing: dataset 2.

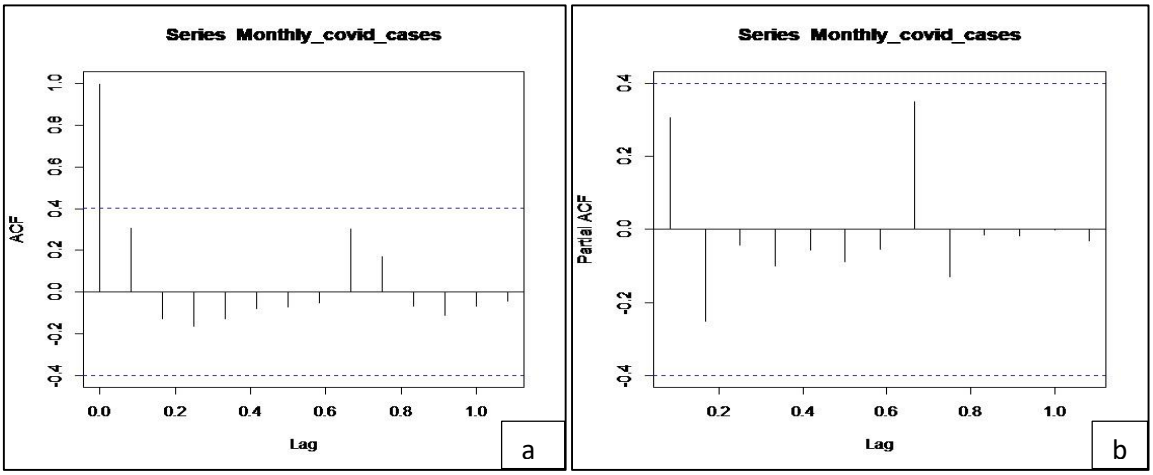

Figure 5(a) and Figure 5(b) are ACF and PACF plots: dataset 1.

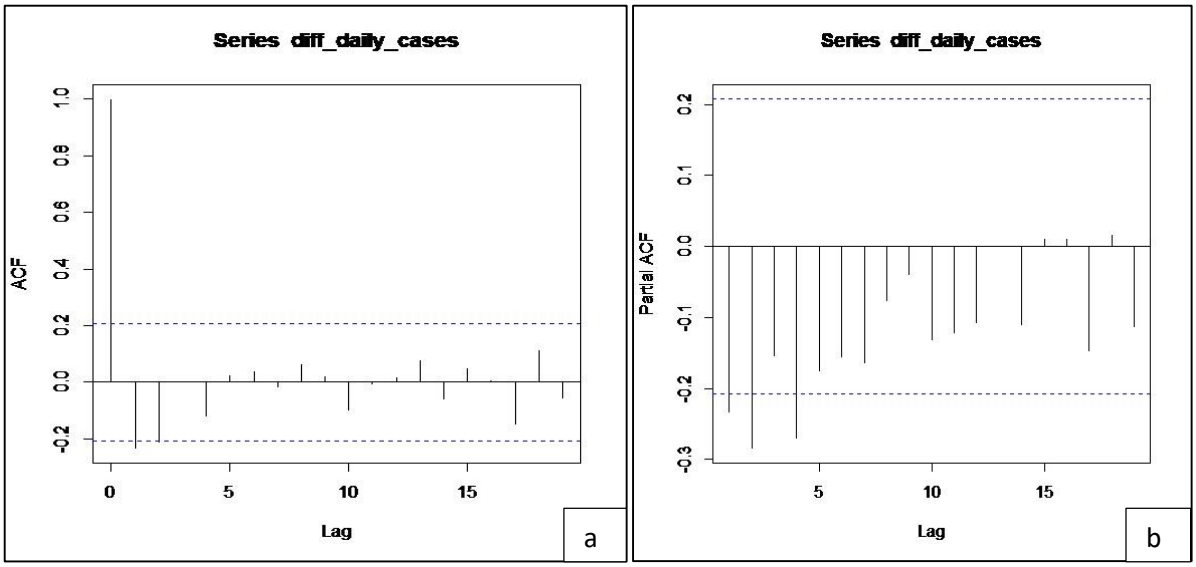

Figure 6(a) and Figure 6(b) are ACF and PACF plots: dataset 2.

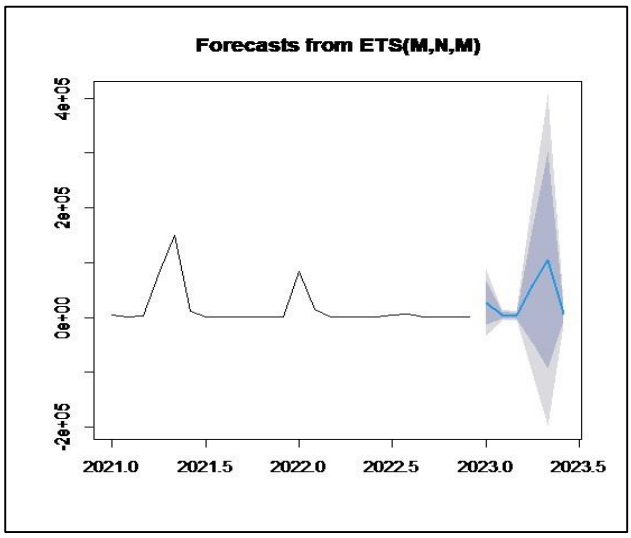

Figure 7: Forecasts result for next six months using single exponential smoothing (dataset 1).

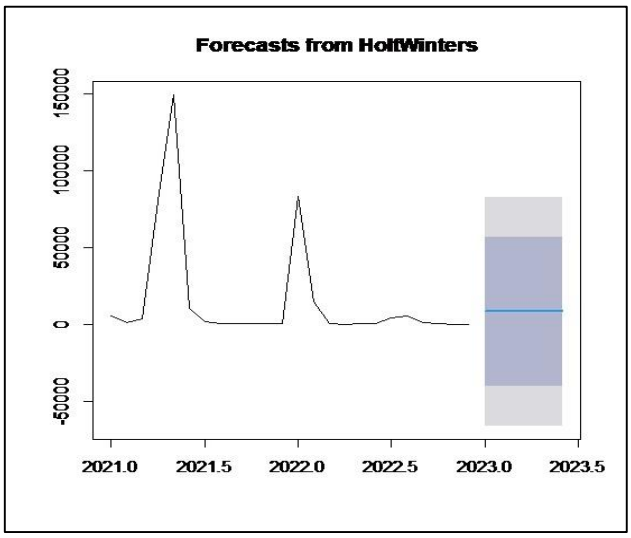

Figure 8: Forecasts result for next six months using ARIMA (0,0,1) (dataset 1).

ARIMA modelling

The partial autocorrelation (PACF) and autocorrelation (ACF) and are two time series plots to identify the behaviour of time series or detecting seasonality as depicted in (Figure 5 (a)) and (Figure 5 (b)) for 'dataset 1', while (Figure 6 (a)) and (Figure 6 (b)) for 'dataset 2'. The Single exponential smoothing and ARIMA were used to forecast the confirmed cases in upcoming months (Table 4), it was seen that COVID-19 cases to peak in May 2023, if the virus continue to follow the same trend as in the past. However the situation will be better further, as depicted in (Figure 7) and (Figure 8). Further based on likelihood function, Akaike information criteria (AIC) and Bayesian information criteria (BIC) were calculated and lower values of these parameters described ARIMA (0,0,2) as the best model which was fitted on the most recent data (dataset 2) as represented in (Table 5).

(Table 6) shows the performance measures of different time series models. The smaller values of mean absolute error (MAE) and root mean square error (RMSE), yielded the better fit. The predictive models for short term duration (dataset 2) yielded better accuracy in terms of smaller values of measures as compared to data recorded for longer duration (dataset 1), suggesting time series models best to deal with short term forecasting. The ARIMA (0,0,2) emerged out to be more accurate than exponential smoothing model while employed on 'dataset 2'. Triple exponential smoothing (or, Holt-Winters) model is best when dealing with 'dataset 1' for the longer period.

DISCUSSION

COVID-19 pandemic across the world has influenced way of living of individuals and health sector is one of the most affected domain. Sarbhan et al. (2020) revealed that ARIMA (0,1,0) performed the best model with MAPE value of 16.01 and BIC value of 4.17 in prediction of COVID-19 cases in Malaysia.[20] Dealing with medical data is a tedious task, especially when working on forecasting COVID-19 data as the virus and its variants have shown a dramatically change over the period. [21]

Different models have been introduced to forecast pandemic outbreak since it's arise. A study compared various time series models to ten different countries while estimating the proportion of active cases in the overall population.[22] The present study is an attempt to compare exponential smoothing and ARIMA time series models for forecasting COVID-19 cases on the basis of performance measures.

In a study conducted by Djakaria et al. (2020), Holt-Winter exponential smoothing was an ideal model for predicting COVID-19 cases in Indonesia, with parameters (α=0.1) and (Y=0.5) and smaller value of MAPE.[23] Our study is also in align with this as it also indicates that Holt Winters smoothing is best for disease seasonality predictions and ARIMA model (0,0,2) seemed to be more accurate for recent data (dataset 2). The study findings are nearly similar to a study by Elsmih et al. (2020) that recommended ARIMA instead of Holt's smoothing for predicting COVID-19 cases in Sudan based on BIC and AIC measurements.[18] Another study depicted the same by concluding that ARIMA (8,2,1) emerged out best models amongst four chosen models: ARIMA, Exponential smoothing, SARIMA, and SARIMAX based on RMSE.[24]

The positive, deceased, and cured cases of COVID-19 were the significant predictors for daily active cases.[25] A study conducted in 2020, also used partial autocorrelation (PACF) correlogram and autocorrelation function (ACF) to estimate the ARIMA model's parameters.[26] To assess how seasonality affected the forecast, they used a logarithmic transformation and their correlogram reflecting the ACF and PACF revealed that seasonality has no bearing on the prevalence or incidence of COVID-19, which was in line with our study as well. However, in the current study Holt winters smoothing did show seasonal effect on the cases.

Only one model cannot predict the definite trend of COVID-19 cases with rest to international travel, virus mutation, population density, human behaviors, etc. Our study findings were in relevant with a study conducted in Saudi Arabia that used the Prophet and ARIMA models for forecasting daily cases, recovery cases, and fatalities for COVID-19.[27] They claimed that neither model was well suited to deal with such data trends. They occasionally performed worse than the baseline model in terms of RMSE and MAE. To increase the depth of the experiment, Multivariate forecasting can be used in order to get the best results because more data sources can increase accuracy. Hybrid model applied to time series analysis gives better results.

Similarly, in a comparative study conducted in Chile, ARIMA and Damped Trend technique were found to be more accurate for predicting COVID-19 cases and deaths respectively.[28] The present study indicated ARIMA (0,0,2) to be more accurate predictive model to deal with data collected for short duration of last three months (dataset 2).

Limitations

The limitation of the study can be considered in terms of changing behaviour of virus continuously and undefined trends of pandemic over time, as both the factors can influence the forecasting results and models accuracy.

CONCLUSION

The present research study established the effectiveness of exponential smoothing model and ARIMA on the basis of performance measures, separately for past two years monthly time series based data and recently obtained data

on confirmed cases. These predictive models on recent dataset is more accurate in terms of smaller values of performance measures as compared to that of longer period. Forecasting of COVID-19 cases play a crucial role for governments and health professionals in framing policies which would result in prevention of the spread of pandemic in future and to make necessary steps for public health care.

ACKNOWLEDGEMENTS

The authors express their gratitude to Swami Rama Himalayan University for their support, also they thank Dr. Akanksha Verma, a post graduate student of Department of Physiotherapy, Himalayan Institute of Medical Sciences, Dehradun, for her help during initial phase of data processing.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun. 2020;109:102433.
2. Khan FM, Gupta R. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. J Safety Science Resilience. 2020;1(1):12-8.
3. Kurbalija V, Radovanovi´c M, Ivanovi´c M, Schmidt D, Trzebiatowski GL, Burkhard HD, Hinrichs C. Time-series analysis in the medical domain: a study of tacrolimus administration and influence on kidney graft function. Comput Biol Med. 2014;50:19-31.
4. Nsoesie E, Beckman R, Shashaani S, Nagaraj K, Marathe M. A simulation optimization approach to epidemic forecasting. PLoS ONE. 2013;8:e67164.
5. Orbann C, Sattenspiel L, Miller E, Dimka J. Defining epidemics in computer simulation models: How do definitions influence conclusions? Epidemics. 2017;19:24-32.
6. Thomson MC, Molesworth AM, Djingarey MH, Yameogo KR, Belanger F, Cuevas LE. Potential of environmental models to predict meningitis epidemics in Africa. Trop Med Int Health. 2006;11:781-8.
7. Liu Q, Li Z, Ji Y, Martinez L, Zia UH, Javaid A, Lu W, Wang J. Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu province of China using advanced statistical time-series analyses. Infect Drug Resist. 2019;12:2311-22.
8. Zhang X, Liu Y, Yang M, Zhang T, Young A, Li X. Comparative study of four time series methods in forecasting typhoid fever incidence in China. PLoS ONE. 2013;8:e63116.
9. Wang Y, Shen Z, Jiang Y. Comparison of ARIMA and GM (1,1) models for prediction of hepatitis B in China. PLoS ONE. 2018;13:e0201987.
10. Zhang L, Wang L, Zheng Y, Wang K, Zhang X, Zheng Y. Time prediction models for echinococcosis based on gray system theory and epidemic dynamics. Int J Environ Res. Public Health. 2017;14:262.
11. Cao L, Liu H, Li J, Yin X, Duan Y, Wang J. Relationship of meteorological factors and human brucellosis in Hebei province, China. Sci Environ. 2020;703:135491.
12. Imai C, Hashizume M. A systematic review of methodology: Time series regression analysis for environmental factors and infectious diseases. Trop Med Health. 2015;43:1-9.
13. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief. 2020;29:1-4.
14. Dehesh T, Mardani HA, Dehesh P. Forecasting of COVID-19 confirmed cases in different countries with ARIMA models. Preprints. 2020;10:1101-45.
15. Chintalapudi N, Battineni G, Amenta F. COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. J Microbiol Immunol Infect. 2020;53:396-403.
16. Singh S, Murali Sundram B, Rajendran K, Boon Law K, Aris T, Ibrahim H, Chandra Dass S, Singh Gill B. Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. J Infect Dev Ctries. 2020;14(9):971-6.
17. Singh S, Sundram BM, Rajendran K, Law KB, Aris T, Ibrahim H, Dass SC, Gill BS. Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. J Infection Developing Countries. 2020;14(09):971-6.
18. Bezerra AK, Santos ÉM. Prediction the daily number of confirmed cases of COVID-19 in Sudan with ARIMA and Holt Winter exponential smoothing. Int J Development Res. 2020;10(08):39408-13.
19. Daily COVID-19 Health Bulletin: Ministry of Health and Family Welfare (MoHFW), Government of Uttarakhand. Available at https:// health. uk.gov.in. Accessed on 12 January 2023
20. Singh S, Sundram BM, Rajendran K, Law KB, Aris T, Ibrahim H, Dass SC, Gill BS. Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. J Infection Developing Countries. 2020;14(09):971-6.
21. Semwal J, Bahuguna A, Sharma N, Dikshit RK, Bijalwan R, Augustine P. Time series analysis of COVID-19 data-a study from Northern India. Indian J Community Health. 2022;34(2):87-9.
22. Papastefanopoulos V, Linardatos P, Kotsiantis S. COVID-19: a comparison of time series methods to forecast percentage of active cases per population. Applied Sci. 2020;10(11):3880.

23. Djakaria I, Saleh SE. COVID-19 forecast using Holt-Winters exponential smoothing. J Physics. 2021;1882(1):012033.
24. Jain A, Sukhdeve T, Gadia H, Sahu SP, Verma S. COVID19 prediction using time series analysis. Int Artificial Intelligence Smart Systems. 2021;25:1599-606.
25. Semwal J, Bahuguna A, Uniyal A, Vyas S. A study to analyse COVID-19 outbreak using multiple linear regression: a supervised machine learning approach. National J Community Med. 2023;14(02):82-9.
26. Gupta A. COVID-19: time series analysis. Int J Scientific Development Res. 2023:2455-631.
27. Zrieq R, Kamel S, Boubaker S, Algahtani FD, Alzain MA, Alshammari F, et al. Time-Series analysis and healthcare implications of COVID-19 pandemic in Saudi Arabia. Healthcare. 2022;10(10):1874.
28. Sandoval BC, Ferreira G, Parra BK, Flores LP. Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques: a comparative study. PLoS One. 2021;16(4):e0245414.

Cite this article as: Bahuguna A, Uniyal A, Sharma N, Semwal J. Comparison of exponential smoothing and ARIMA time series models for forecasting COVID-19 cases: a secondary data analysis. Int J Res Med Sci 2023;11:1727-34.