ORIGINAL ARTICLE

Time Series Analysis of COVID-19 Data- A study from Northern India

Jayanti Semwal[1], Abhinav Bahuguna[2], Neha Sharma[3], Rajiv Kumar Dikshit[4], Rajeev Bijalwan[5], Piyush Augustine[6]

[1]Professor, Department of Community Medicine, Himalayan Institute of Medical Sciences (HIMS), Swami Rama Himalayan University (SRHU), Dehradun; [2]Lecturer, Department of Biostatistics, Himalayan Institute of Medical Sciences (HIMS), Swami Rama Himalayan University (SRHU), Dehradun; [3]Assistant Professor, Department of Community Medicine, Himalayan Institute of Medical Sciences (HIMS), Swami Rama Himalayan University (SRHU), Dehradun; [4]District Surveillance Officer-IDSP(District Nodal officer- Covid-19), Dehradun; [5]Deputy Director, Rural Development Institute (RDI), Swami Rama Himalayan University (SRHU), Dehradun; [6]District Epidemiologist IDSP- cell, Dehradun

Abstract | Introduction | Methodology | Results | Conclusion | References | Citation | Tables / Figures

Corresponding Author

Dr. Neha Sharma, Assistant Professor, Department of Community Medicine, Himalayan Institute of Medical Sciences (HIMS), Swami Rama Himalayan University (SRHU), Dehradun.
E Mail ID: dr_neha2402@yahoo.com

Citation

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 Comm Health. 2022;34(2):202-206. https://doi.org/10.47203/IJCH.2022.v34i02.012

Source of Funding: Nil **Conflict of Interest:** None declared

Received: 05/06/2022; **Revision:** 15/06/2022; **Accepted:** 26/06/2022; **Published:** 30/06/2022

This work is licensed under a Creative Commons Attribution 4.0 International License.

Abstract

The continuing new Coronavirus (COVID-19) pandemic has caused millions of infections and thousands of fatalities globally. Identification of potential infection cases and the rate of virus propagation is crucial for early healthcare service planning to prevent fatalities. The research community is faced with the analytical and difficult real-world task of accurately predicting the spread of COVID-19. We obtained COVID-19 temporal data from District Surveillance Officer IDSP, Dehradun cum District Nodal Officer- Covid-19 under CMO, Department of Medical Health and Family Welfare, Government of Uttarakhand State, India, for the period, March 17, 2020, to May 6, 2022, and applied single exponential method forecasting model to estimate the COVID-19 outbreak's future course. The root relative squared error, root mean square error, mean absolute percentage error, and mean absolute error were used to assess the model's effectiveness. According to our prediction, 5438 people are subjected to hospitalization by September 2022, assuming that COVID cases will increase in the future and take on a lethal variety, as was the case with the second wave. The outcomes of the forecasting can be utilized by the government to devise strategies to stop the virus's spread.

Keywords

COVID-19; Forecasting; Hospitalization; Single exponential smoothing

Introduction

COVID-19 (SARS-CoV-2) outbreak started in the Wuhan region, and it has since spread to other parts of the world, causing a global pandemic. (1) By Dec 31, 2020, COVID-19 had killed over 1.8 million people and infected over 82 million people. (2) Since its initial outbreak, it has spread to 228 countries and territories. (3) In addition to causing significant economic disruptions, this virus has inflicted extreme suffering in every nation. A number of precautionary measures have been taken by the afflicted countries. (4) To halt the spread of the disease, countries have been effectively adopting numerous activities. (5)

It is a disease with specific growth patterns which are non-linear and dynamic in character as the cases may vary depending on the seasons, population, and other factors. (4) COVID-19 predictive analysis has been a major research topic in order to help the government plan for and prevent the development of this disease. (6) Modeling and anticipating the virus's daily spread behavior can healthcare organizations prepare for the expected influx of patients. (7) Several research on projecting the number of COVID-19 cases have recently been published using different models. However, no universal strategy for selecting models for forecasting COVID-19 spread has been established. (8)

Aims & Objectives

To forecast the hospitalization of COVID-19 patients in Uttarakhand using available district Dehradun data from

March 17, 2020, to May 6, 2022, and display the patterns of the disease for the next eight weeks.

Material & Methods

This study utilized secondary data collected from the District Surveillance Officer (Integrated Disease Surveillance cell of district, Dehradun) cum District Nodal Officer for Covid-19 under CMO, Department of Health and Family Welfare, Government of Uttarakhand State, India, to provide the COVID-19 data of district Dehradun for the time period between March 17, 2020 to May 6, 2022 [Letter Reference No.- DSO/IDSP/2021-22/1776A]. With due permission from the health authorities, the data were accessed, analyzed, and displayed the patterns of the disease time series for the next eight weeks. The ethical approval was also taken from the University for utilizing the Government data publishing the results. As per the data records, the patients who were COVID-19 positive and admitted to any hospital in the district Dehradun (complete enumeration) were taken in the study.

A time series is a group of data points that have been arranged chronologically or listed, enumerated, or graphed. The data can be compared to random sample data that have additional information that can be extracted because it is organized using somewhat deterministic timestamps. Over the past few decades, time series forecasting models have seen a great deal of growth and improvement. TBAT, Prophet, ARIMA, Moving Average, Neural Basis Expansion Analysis (N-BEATS), Single Exponential Method, and Double Exponential Method are some of the forecasting techniques.

(i) This study utilizes the Single exponential smoothing (SES) technique which is used for forecasting when dealing with data distributed according to time with no clear trend and seasonal pattern. The general equation for single smoothing is described as:

$S_t = \alpha\, x_t + (1 - \alpha)\, S_{t-1}$

S_t = smoothed statistic

$x_{t=}$ the actual value in time period t

α= smoothing factor, usually lies between 0 to 1

(ii) To evaluate our Forecast model accurately, the following statistic is calculated:

Further mean absolute percent error (MAPE) used to assess the performance of our forecast model. Less the value of MAPE, better is the forecast model.

Root mean square error (RMSE) is calculated as,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean absolute error (MSE) is calculated as,

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

The mean absolute percent error (MAPE) is calculated as,

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Lesser the value of above statistics will yield better fitted forecast model. y_i will be treated as actual output, whereas $\hat{y}_i$ is predicted output in above defined statistics.

(iii) To measure the uncertainty associated with our forecasted value, prediction interval will be calculated with the following formula:

$$Prediction\ Interval = Predicted\ Value \pm z\sqrt{MSE}$$

The value of z is 1.96 at 95 % level of prediction level.

Statistical Analysis: The data were entered in MS excel 2010. Statistical analysis was performed using R software version 4.1.3 and SPSS version 22 and. Descriptive statistics were presented in the form of frequency tables. To check the stationarity of the data, some statistical tests were performed. Augmented Dickey Fuller Test was implied where following hypothesis was set up:

H_0: the hospitalization data of patients are not stationary

H_1: the hospitalization data of patients are stationary

To check the existence trend in the dataset, another test was used to check the stationarity of data i.e., Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS), where the hypothesis was as follows:

H_0: the hospitalization of patients are stationary

H_1: the hospitalization of patients are not stationary

For making the data stationary, differencing method was used, which shows a series of difference that how one period changes to the next. It is a difference between successive observations ($Y_t'=Y_t-Y_{t-1}$). After applying differencing, tests for stationarity were again implied.

Results

The analysis was conducted on a sample of 1, 45,626 patients using available data from March 17, 2020, to May 6, 2022 (till the latest update, however, the status of newly admitted patients is not updated till the analysis). Maximum number of COVID-19 patients from our sample belonged to the middle age group i.e., 18-39 years (50.2%) followed by 40-59 years (28.5%). Table 1 shows that out of the said registered COVID-19 patients, the majority of patients (97.6%) were discharged. A total of 1.7% of deaths due to COVID-19 was recorded in the state of Uttarakhand during the above-mentioned period.

Table 2 shows that the highest mortality was found in the 60 + year's age-group (6.1%), followed by the age group of 40-59 years with 1.9% death registered. However, there is not much difference in the recovery rate amongst different age groups but to an extent age group 0-17 years has registered the highest recovery rate (more than 99%). Overall, the recovery rate was good for all the age-group patients except for the 60+ years, whose recovery rate was the lowest i.e., 93.1%.

Figure 1 shows that the maximum number of hospitalizations amongst COVID-19 patients was in April-May 2021. The hospitalization rate again increased in January 2022 but started declining thereafter.

Table 3.1 shows that on performing Dickey-Fuller test the p-value was greater than 0.1, hence the data was non-stationary. While, for KPSS test the , p-value was ≥ 0.1, so the data was stationary at a 10% level of significance.

Figure 2 shows the time series plot after differencing to make the series of hospitalization stationary. After applying first order differencing, the data was again tested for stationarity using Dickey-Fuller and KPSS test. Table 3.2 showed that in the Dickey Data test p-value was less than 0.01, making the data stationary. Also, for the KPSS test, the p-value came out to be ≥ 0.1, so the data was stationary at 10 % level of significance.

Due to the rise in COVID-19 cases without any trend and seasonal pattern, a single exponential smoothing model was used for forecasting. The calculated value is of smoothing factor (α) I s 0.1 0.(Table 4)

Figure 3 shows that if the trend of COVID-19 behavior remains the same (from mid-march 2020 to April 2022), there might be a peak of cases in Sept 202 2. To further validate the forecasting the data was summarized again, and time series plot to observe trends and patterns from September 2021 to April 2022 was performed assuming that the series will follow the seasonality pattern in the future. If so, the peak would have been in recent times and cases will fall down in the future (Figure 4).

In the forecasted model, the R-square value was 67 % and the mean absolute percentage error (MAPE) was approx. 10% that describes the model is well fitted. (Table 5)

Discussion

Future prediction through forecasting estimates is what will be the result in the upcoming future based on the ongoing trends and it's quite interesting as we get to know the future possibility which helps us in planning policies and taking decisions. When it comes to the prediction of medical data it becomes a challenging task, especially when dealing with COVID- 19 data, as the pandemic had different phases from March 2020 to date. The population has been affected by different kinds of COVID variants, such as the deadly known delta variant, which was responsible for the second wave of covid-19 in India, that resulted in the maximum number of deaths and hospitalization, followed by the omicron variant and so on. The central, as well as state governments, have provided guidelines to control the spread of pandemic from time to time, such as the imposition of lockdown, night curfew, travel, and tourism restrictions from another state to Uttarakhand and vice-versa, restriction of people on any kind of gathering or on occasions (a limited persons, marriages, online education, opening of gyms, cinemas, parks with only 50 % capacity, etc.). Despite so many efforts from the central and state government and also local administrators it was difficult to get control over the pandemic completely as it is a basic need of an individual to go to his/her workplace, especially for those who are totally dependent on daily wages and more importantly to boost the up economy of the country.

So, the present study is an attempt to forecast the number of patients who will be subjected to hospitalization based on the patients admitted to the hospital from mid-March 2020 to April 2022. This is also one of the limitations of our study because it will be true only if the number of cases will increase in the future with the same pattern as in the past and also on the severity of COVID- 19 variants. In the future, researchers can use COVID-19 to investigate prediction models like artificial neural networks (ANN), Bayesian networks, and Support Vector Machines (SVM). This algorithm can also be used to anticipate future pandemics and any form of the disease that affects humans.

Conclusion

The prediction of our study states that there will be 5438 patients who are believed to be hospitalized by September 2022 assuming that COVID- 19 cases will rise in the future with a deadly variant as was the case with the second wave. Only providing a value of a forecast (5438 hospitalization in Sep. 2022) is always not enough as the hospitalization of patients will depend on both the number of COVID- 19 cases the and severity of the COVID-19 variants in upcoming days. Because of this uncertainty, we have calculated the prediction interval at 95% prediction level which is (0, 29380.65), which tells the uncertainty associated with our forecasts. We set the lower value of the prediction interval to 0 as a negative forecast value does not sound good mathematically. However, due to various factors under consideration such as a large number of populations have taken both the doses of COVID- 19 vaccines, vaccine boosters, a decline in COVID- 19 cases in recent time, formation of anti-bodies (as most of the people already have affected with COVID- 19), government policies and awareness of people will have a positive effect on the pandemic in terms of recovery rate and the spread of cases. The suggested model may be helpful in forecasting future cases of infection in the current scenario, provided that the pattern of virus dissemination does not change abnormally. The estimated prediction interval (-18504.64, 29380.65) makes it clear that the likelihood of hospitalization due to COVID will be lower than what we have predicted, but there is still a lot of uncertainty. (1) The model proved to be the most effective model for short-term forecasting in the time series and the outcome in the very near months, in contrast to other models, which call for a significant number of previously calculated timestamp samples. Future improvements to the models' predictive accuracy will include the development of ensembles of the ones that have been presented, which combine the best aspects of each model to lower overall error, as well as the use of multivariate time series modeling to take other factors into account that may be

directly or indirectly related to the pandemic's spread. Future plans include using some form of transfer learning to disseminate knowledge from one nation to another in an effort to identify the main contributing variables to the true origin of the spread. (9)

Recommendation

The use of artificial neural networks (ANN) and Bayesian networks as prediction models in COVID-19 is something that researchers can study in the future.

Limitation of the study

The data set used for the forecast was rather small, and the prediction was made in the context of a pandemic with significant levels of data set fluctuations. The output would have been more accurate if the dataset were more comprehensive and had lower variation. Moreover, the forecasting used is based on the hospitalization of patients instead of the number of cases (or positive cases). The vaccination rate also posed a significant impact on the precision of our model.

Relevance of the study

Modeling and anticipating the virus's daily spread behavior can help healthcare organizations prepare for the expected influx of patients. Many types of research on forecasting the number of COVID-19 cases have been published using the ARIMA model, Holt's linear trend model, and the SIR state transition model but no universal model has been established. A state-wise model can be framed using the available data and checking the trends of disease as it could have an impact on government policies, containment rules, the health system, and social life.

Authors Contribution

JS: Conception and design, article drafting, final approval of the version to be published, AS: Analysis and interpretation of data, article drafting and final approval of the version to be published, NS: Analysis and interpretation of data, article drafting and final approval of the version to be published, RKD: Conception and design, acquisition of data, revising the manuscript critically for important intellectual content, RB: Conception and design, revising the manuscript critically for important intellectual content PA: Conception and design, revising the manuscript critically for important intellectual content.

Acknowledgment

The authors extend their heartfelt gratitude to the District Surveillance Officer of the Integrated Disease Surveillance cell of the district, Dehradun (Uttarakhand) under CMO, Department of Health and Family Welfare, Government of Uttarakhand State, India, the for smooth availability of data and permitting us to work on it.

References

1. Singh S, Chowdhury C, Panja AK, Neogy S. Time Series Analysis of COVID-19 Data to Study the Effect of Lockdown and Unlock in India. J Inst Eng Ser B. 2021;102(6):1275–81. Available from: https://link.springer.com/article/10.1007/s40031-021-00585-7
2. The impact of COVID-19 on global health goals. [Accessed on 25.06.2022]. Available from: https://www.who.int/news-room/spotlight/the-impact-of-covid-19-on-global-health-goals
3. Countries where Coronavirus has spread - Worldometer. [Accessed on 25.06.2022]. Available from: https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/
4. Bodapati S, Bandarupally H, Trupthi M. COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks. 2020 IEEE 5th Int Conf Comput Commun Autom ICCCA 2020. 2020 Oct 30;525–30. [Accessed on 25.06.2022]. Available from: doi: 10.1109/ICCCA49541.2020.9250863.
5. Satrio C.B.A, Darmawan W, Nadia BU, Hanafiah N. Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. Procedia Computer Science. 2021;179:524–32.
6. Murray CJ. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv. 2020;2020.03.27.20043752.: doi.org/10.1101/2020.03.27.20043752
7. Kumar N, Susan S. COVID-19 Pandemic Prediction using Time Series Forecasting Models. 2020 11th Int Conf Comput Commun Netw Technol ICCCNT 2020. 2020 Jul 22 [Accessed on 25.06.2022]; Available from: doi: 10.1109/ICCCNT49239.2020.9225319.
8. Abotaleb M, Makarovskikh T, Rojas F, Herrera LJ, Pomare H. System for Forecasting COVID-19 Cases Using Time-Series and Neural Networks Models. Eng Proc 2021;5(1):46. [Accessed on 25.06.2022]. Available from: https: https://doi.org/10.3390/engproc2021005046 .
9. Chyon FA, Suman MN, Fahim MR, Ahmmed MS. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. Journal of Virological Methods. 2022;301:114433.

Tables

TABLE 1 FREQUENCY TABLE SHOWING OUTCOME STATUS OF THE COVID-19 PATIENTS

Outcome status of the COVID-19 patients	Frequency	Percent
Active	75	0.1
Death	2459	1.7
Discharge	142114	97.6
Migrated	967	0.6
Not Updated	8	0.0*
Referred	3	0.0*
Total	145626	100.0

*The values are considered approximately equal to 0.

TABLE 2 FREQUENCY TABLE SHOWING AGE GROUP WISE OUTCOME STATUS OF THE COVID-19 PATIENTS

Age group wise Outcome Status		Outcome Status						Total
		Active n (%)	Death n (%)	Discharge n (%)	Migrated n (%)	Not Updated n (%)	Referred n (%)	
Age group	< 18 years	20 (0.2)	19 (0.2)	9214 **(99.1)**	44 (0.5)	2 (0.0)*	0 (0.0)*	9299
	18-39 years	24 (0.0)*	274 (0.4)	68656 (98.9)	520 (0.7)	2 (0.0)*	1 (0.0)*	69477
	40-59 years	21 (0.0)*	839 (1.9)	44000 (97.5)	244 (0.6)	1 (0.0)*	2 (0.0)*	45107
	60 years and above	10 (0.0)	1327 **(6.1)**	20244 (93.2)	159 (0.7)	3 (0.0)*	0 (0.0)*	21743
Total		**75 (0.1)**	**2459 (1.7)**	**142114 (97.6)**	**967 (0.6)**	**8 (0.0)***	**3 (0.0)***	**145626**

*The values are considered approximately equal to 0.

TABLE 3.1 TESTS FOR CHECKING THE STATIONARY STATUS OF THE DATA BEFORE DIFFERENCING

Statistics	Value	Dickey-Fuller
Dickey Fuller	-2.7939	0.2699
KPSS	0.10871	0.1

TABLE 3.2: TESTS FOR CHECKING THE STATIONARY STATUS OF THE DATA AFTER DIFFERENCING

Statistics	Value	P-value
Dickey-Fuller	-3.7153	< 0.01
KPSS	0.04177	≥ 0.1

TABLE 4 FORECASTING FOR THE MONTH OF SEPTEMBER 2022

Month	Forecast	95% Prediction level	
		Lower	Upper
September 2022	5438.009	0	29380.65

TABLE 5 FORECASTED MODEL STATISTICS

Statistic	Model fitting
R-Squared	67%
RMSE	73.8%
MAPE	10.44%
MAE	45.19%

Figures

FIGURE 1 COVID 19 HOSPITALIZATION PATTERN FROM MID MARCH 2020 TO MAY 2022 (FIRST WEEK)

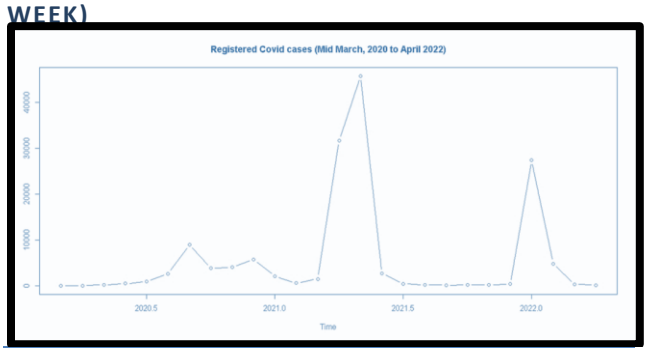

FIGURE 2 GRAPH SHOWING DIFFERENCING METHOD FOR MAKING DATA STATIONARY

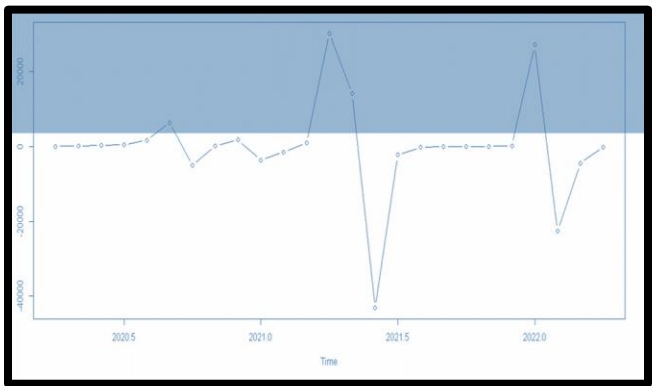

FIGURE 3 : FITTED GRAPH TO OBTAIN FORECAST VALUE USING SINGLE EXPONENTIAL SMOOTHING

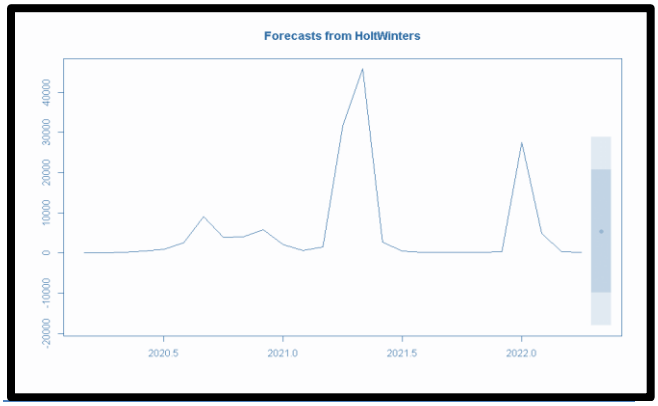

FIGURE 4 FITTED VS. FORECASTED GRAPH CONSIDERING THE SEASONALITY PATTERN