TWO YEARS AFTER THE PANDEMIC, FORECASTING THE NUMBER OF NEW COVID-19 CASES IN THE UNITED STATES OF AMERICA

N Sunengsih[1], IGNM Jaya[1*]
[1]Department of Statistics, Padjadjaran University, Bandung, Indonesia
*Corresponding author: mindra@unpad.ac.id

Abstract

United States of America is one of the countries in the world with the highest number of COVID-19 incidences. After two years of the pandemic, COVID-19 infections in USA remain uncontrolled. Due to the omicron variant's rapid transmission, the daily number of cases in America has become more out of control. The vaccine program has failed to protect America from the highest daily rate of infection in the world. The purpose of this study was to forecast the daily incidence of COVID-19 from January to March 2022. We used a machine learning method called extreme gradient boosting (XGBOOT), as well as a more traditional approach called seasonal autoregressive moving average (ARIMA). Both models estimate that the daily number of cases in America will continue to rise until March. According to model comparisons, the XGBOOTS model outperforms the SARIMA model in terms of forecasting performance.

Keywords: *COVID-19, USA, XGBOOTS, SARIMA*

1. INTRODUCTION

Since its discovery in Wuhan, China, two years ago, COVID-19 has spread to over 200 countries, with severe health, socioeconomic, and political consequences (1, 2). According to the WHO, the overall number of COVID-19 cases has surpassed 328 million as of January 17, 2022, with a mortality toll of more than 5.5 million [3]. This figure continues to rise, with each new wave of COVID-19 increasing at a faster rate than the prior waves of COVID-10.

The United States is the primary source of COVID-19 cases worldwide (3, 4). With a population of over 333 million people, or 4.20 % of the world's total population, the number of COVID-19 cases was over 64.7 million as of January 17, 2022, and the death toll exceeded 800,000 (3). Since November 26, 2021, when the WHO first stated it, the omicron variation has been designated the primary cause of the high number of illnesses in America. In comparison to the delta variant, this variant has a significantly high infection rate (5). Extra measures must be undertaken to bring the large number of cases in America under control.

Objective

The purpose of this study is to forecast the number of COVID-19 cases in the United States of America between January 18, 2022 and March 28, 2022. The forecasted results are likely to provide the administration with a foothold in reducing the growing number of COVID-19 cases. There are numerous forecasting methods, including traditional approaches such as SARIMA and machine learning methods such as XGBOOST. We compare the predicting findings from two different methodologies in order to arrive at a more precise predicted value.

Hypothesis

The number of new cases in the United States of America is expected to be reasonably high for the following two months 18 January 2022 to 18 March 2022

2. MATERIAL AND METHOD

2.2.1. SARIMA

Box-Jenkins ARIMA is the most widely used technique for forecasting ordinary time series data, including stationary and non-stationary. The objective is to develop an Autoregressive Integrated Moving Average (ARIMA) model that accurately depicts the process of data generation. The fundamental Box-Jenkins expression is as follows (6, 7):

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j e_{t-j}, i = 1, \dots, p \text{ and j} = 0,1, \dots, q \quad (1)$$

where y_t is a stationary stochastic process with a non-zero mean, α_0 denotes the intercept of the ARIMA model, and e_t denotes error term. Then, the i and j terms represent the model's autoregressive and moving average components, respectively [8]. For seasonal models, Equation 1 can be expanded to become $SSARIMA(p,d,q)(P,D,Q)[s]$. Where P,D,Q are equivalent to p,d,q but are used in this context as seasonal components with s denoting seasonal order (8).

2.2.4. Extreme Gradient Boosting (XGB)

Extreme gradient boosting is one of the most loved gradient bosting algorithm. It can be used for regression and classification and it is constructed based on gradient boosting framework. Suppose we have a data set $D = \{x_t : y_t\}$ containing T observation, where x denotes training and y is a target variable. XGBOOST assumes there are G additive function of boosting with formulation (9):

$$\hat{y}_t = \sum_{g=1}^{G} f_g(x_t) \quad (2)$$

where $\hat{y}_t$ denotes the forecast for time t at the g -th boost and f_g denotes a tree constructed L with leaf j with a weight score of ω_j. XGBOOST's estimation procedure is based on minimization of the loss function L_g as follows [8] [9]:

$$L_g = \sum_{t=1}^{T} l(y_t, \hat{y}_t) + \gamma T + \frac{1}{2}\lambda\|\omega\| \quad (3)$$

where l is a differentiable convex loss function that measure the different between prediction $\hat{y}_t$ and target value y_t.

2.3. Forecast evaluation methods

To compare the performance of extreme gradient boosting and SARIMA in forecasting new COVID-19 cases on a daily basis, we employ Mean Absolute Error and Root Mean Square Error, as defined below (10):

$$MAE = \frac{1}{H}\sum_{h=1}^{H}(y_{t+h} - \hat{y}_{t+h}) \quad (10)$$

where y_{t+h} denotes data testing at $t+h$ period and $\hat{y}_{t+h}$ is the forecast value for $t+h$ period with H is length of forecast.

$$RMSE = \sqrt{\frac{1}{H-1}\sum_{h=1}^{H}(y_{t+h} - \hat{y}_{t+h})^2} \quad (11)$$

The model with the fewest MAE and RMSE values is the best model.

3. RESULTS AND DISCUSSION

COVID-19 incidences totaled 66,529,283 in the United States of America throughout the study period, accounting for 20.25 % infection worldwide. The data were accessed from

https://github.com/CSSEGISandData/COVID-19. The daily and cumulative temporal pattern of COVID-19 incidences in USA is presented in Figure 1

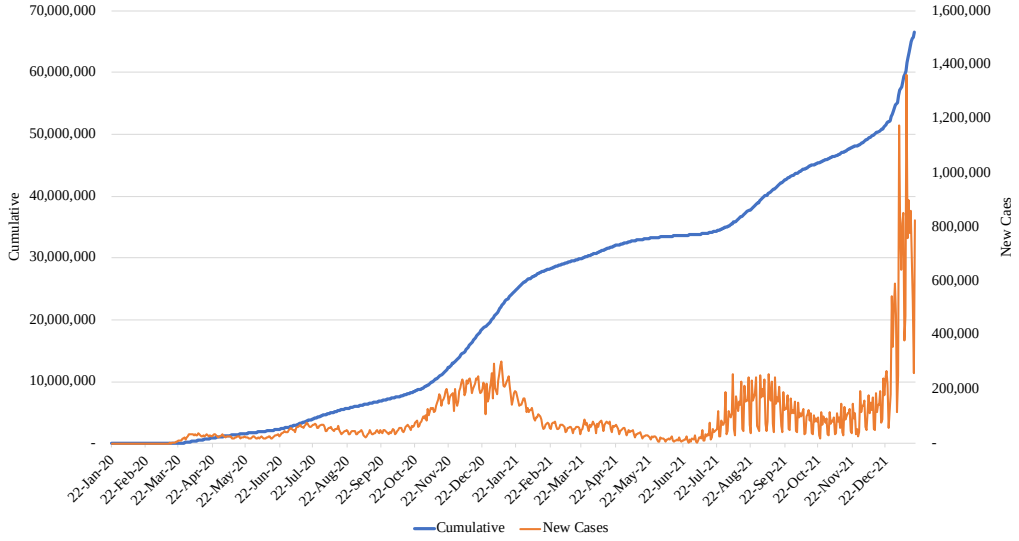

Figure 1 The daily and cumulative temporal pattern of COVID-19 incidences in USA during 22 January 2020 to 17 January 2022

Between January 22, 2020 and January 17, 2022, the USA appears to have seen six waves of COVID-19 infections. The first wave occurs in March 2020; the second wave occurs in May 2020; the third wave occurs in December 2020; the fourth wave occurs in March 2021; the fifth wave occurs in August 2021; and the sixth wave occurs in December 2022. The sixth wave is the most severe ailment, with an average daily incidence of 800,000 per day.

To avoid the worst health and socioeconomic consequences of COVID-19, it is critical to obtain reliable information about the COVID-19 situation over the next several months. To gather reliable information for the next several months, we apply forecasting tools. We used two forecasting techniques: machine learning via extreme gradient boosting (XGBOOST) and classical regression using seasonal autoregressive moving average (ARIMA). Both methods were implemented in R using the packages `forecastxgb` (11) and `forecast` (7) for XGBOOST and SARIMA, respectively.

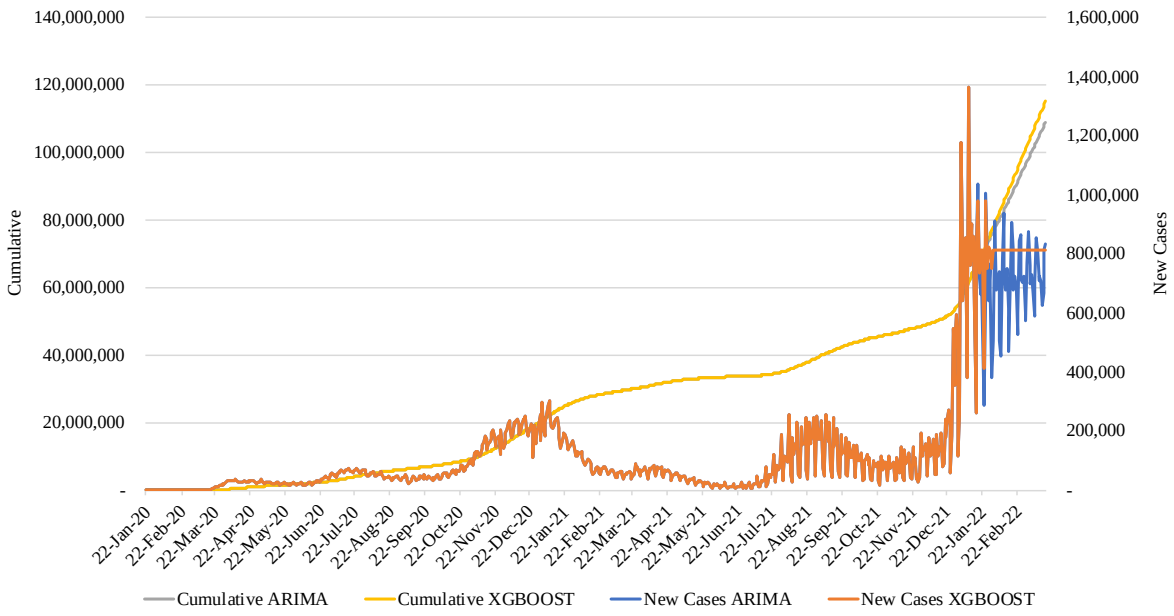

Figure 2. Forecasting results XGBOOST versus SARIMA for period 22 January 2020 to 18 March 2022

Figure 2 shows the comparison of the temporal trend of COVID-19 during period 22 January 2020 to 18 March 2022 between XGBOOST and SARIMA models. The new cases and cumulative cases values of forecast over period 18 January 2022 to 18 March 2022 are presented in Table 1.

Table 1. The new cases and cumulative cases values of forecast over period 18 January 2022 to 18 March 2022 based on XGBOOST and SARIMA models.

Period	New Cases ARIMA	New Cases XGBOOST	Cumulative ARIMA	Cumulative XGBOOST	Period	New Cases ARIMA	New Cases XGBOOST	Cumulative ARIMA	Cumulative XGBOOST
18-Jan-22	1,036,482	977,601	67,565,765	67,506,884	17-Feb-22	817,015	813,742	88,292,746	91,588,504
19-Jan-22	718,455	735,573	68,284,219	68,242,457	18-Feb-22	678,933	813,742	88,971,679	92,402,246
20-Jan-22	662,385	786,900	68,946,604	69,029,358	19-Feb-22	726,252	813,742	89,697,930	93,215,988
21-Jan-22	809,837	813,742	69,756,441	69,843,100	20-Feb-22	688,738	813,742	90,386,668	94,029,731
22-Jan-22	544,891	813,742	70,301,332	70,656,842	21-Feb-22	528,894	813,742	90,915,563	94,843,473
23-Jan-22	286,077	412,620	70,587,409	71,069,462	22-Feb-22	599,231	813,742	91,514,793	95,657,216
24-Jan-22	633,760	945,277	71,221,169	72,014,740	23-Feb-22	843,573	813,742	92,358,366	96,470,958
25-Jan-22	1,004,821	977,601	72,225,989	72,992,341	24-Feb-22	862,853	813,742	93,221,219	97,284,700
26-Jan-22	829,897	748,711	73,055,887	73,741,052	25-Feb-22	713,531	813,742	93,934,750	98,098,443
27-Jan-22	641,198	821,633	73,697,084	74,562,685	26-Feb-22	703,200	813,742	94,637,951	98,912,185
28-Jan-22	763,613	813,742	74,460,697	75,376,427	27-Feb-22	723,632	813,742	95,361,583	99,725,928
29-Jan-22	678,342	813,742	75,139,039	76,190,169	28-Feb-22	601,417	813,742	95,963,000	100,539,670
30-Jan-22	380,781	750,971	75,519,820	76,941,140	01-Mar-22	573,912	813,742	96,536,912	101,353,412
31-Jan-22	505,074	813,742	76,024,894	77,754,883	02-Mar-22	768,444	813,742	97,305,356	102,167,155
01-Feb-22	905,838	813,742	76,930,731	78,568,625	03-Mar-22	874,036	813,742	98,179,391	102,980,897
02-Feb-22	910,392	813,742	77,841,123	79,382,368	04-Mar-22	762,140	813,742	98,941,531	103,794,640
03-Feb-22	678,127	813,742	78,519,250	80,196,110	05-Mar-22	695,921	813,742	99,637,452	104,608,382
04-Feb-22	707,404	813,742	79,226,654	81,009,852	06-Mar-22	729,233	813,742	100,366,685	105,422,124
05-Feb-22	740,964	813,742	79,967,618	81,823,595	07-Mar-22	664,669	813,742	101,031,354	106,235,867
06-Feb-22	503,931	813,742	80,471,549	82,637,337	08-Mar-22	587,498	813,742	101,618,852	107,049,609
07-Feb-22	453,522	813,742	80,925,070	83,451,080	09-Mar-22	704,743	813,742	102,323,595	107,863,352
08-Feb-22	781,548	813,742	81,706,619	84,264,822	10-Mar-22	853,594	813,742	103,177,190	108,677,094
09-Feb-22	936,164	813,742	82,642,783	85,078,564	11-Mar-22	806,542	813,742	103,983,732	109,490,836
10-Feb-22	747,278	813,742	83,390,061	85,892,307	12-Mar-22	709,405	813,742	104,693,136	110,304,579
11-Feb-22	674,933	813,742	84,064,993	86,706,049	13-Mar-22	721,715	813,742	105,414,852	111,118,321
12-Feb-22	747,767	813,742	84,812,761	87,519,792	14-Mar-22	706,953	813,742	106,121,805	111,932,064
13-Feb-22	614,267	813,742	85,427,028	88,333,534	15-Mar-22	624,210	813,742	106,746,015	112,745,806
14-Feb-22	469,917	813,742	85,896,945	89,147,276	16-Mar-22	665,798	813,742	107,411,813	113,559,548
15-Feb-22	670,937	813,742	86,567,881	89,961,019	17-Mar-22	813,795	813,742	108,225,608	114,373,291
16-Feb-22	907,850	813,742	87,475,731	90,774,761	18-Mar-	833,808	813,742	109,059,415	115,187,033

Period	New Cases ARIMA	New Cases XGBOOST	Cumulative ARIMA	Cumulative XGBOOST	Period	New Cases ARIMA	New Cases XGBOOST	Cumulative ARIMA	Cumulative XGBOOST
					22				

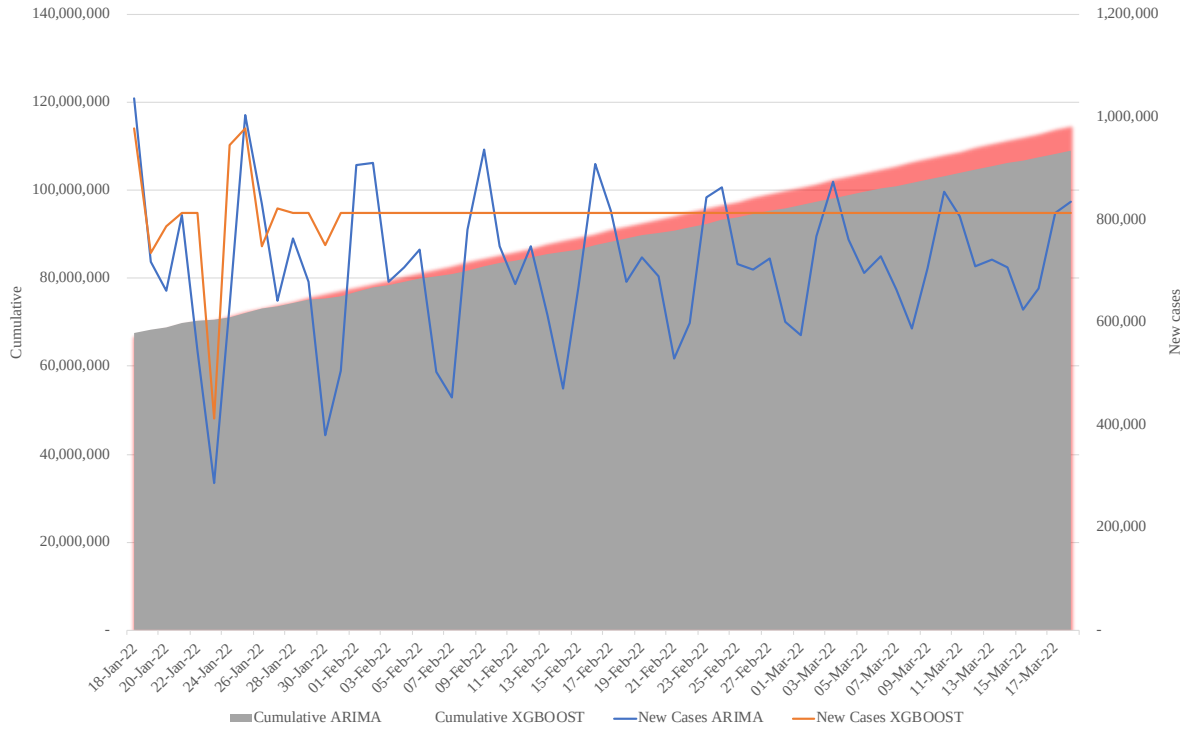

Figure 3. Forecasting results XGBOOST versus SARIMA for period 18 January 2022 to 18 March 2022

Figure 3 presents clearly the comparison of the forecasting results of the XGBOOST and SARIMA models for period 18 January 2022 to 18 March 2022. In comparison to the SARIMA model, the forecasting result of XGBOOST is comparatively stable across the forecasting period. We project that between 18 January and 18 March 2022, there will be 115,187,033 and 10,059,415 new cases, respectively, using the XGBOOST and SARIMA models.

Table 2. Models comparison

	RMSE	MAE
Extreme Gradient Boosting	10713.000	6401.691
SARIMA	44305.530	19409.740

The root mean square error (RMSE) and mean absolute error (MAE) are used to compare the predictive performance of the XGBOOST and SARIMA models. The model with the lowest RMSE and MAE is the best prediction model. As shown in Table 2, the XGBOOST model is the most accurate for projecting COVID-19 in the United States of America.

4. CONCLUSION

America has the largest number of COVID-19 cases and deaths worldwide. After two years and increasing vaccination rates, the number of COVID-19 cases in the United States remains extremely high, despite the possibility of lowering the fatality rate. In January 2022, the number of new cases averaged 800,000 per day. Numerous instances.

We forecasted the number of COVID-19 instances in the United States between January 18, 2022 and the next sixty days in this study. The statistics indicate that the daily case count remains about 800,000 per day. As a developed country with a high level of citizen mobility, the United States must quickly implement policies

to contain the spread of COVID-19 and not rule out the possibility of the United States becoming the global epicenter of COVID-19.

Strictly enforcing the use of masks by all citizens who are not at home can be an effective way to prevent the spread of COVID-19, particularly the Omicron variant.

References

[1] Q. Lin, S. Zhao, D. Gao, Y. Lou, S. Yang, S. S. Musa and et al., "A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action," Int J Infect Dis, vol. 93, pp. 211-216 , 2020.

[2] M. Liu, J. Ning, Y. Du, J. Cao, D. Zhang, J. Wang and M. Chen, "Modelling the evolution trajectory of COVID-19 in Wuhan, China: experience and suggestions," Public Health, vol. 183, pp. 76-80, 2020.

[3] WHO, "WHO Coronavirus (COVID-19) Dashboard," 17 January 2022. [Online]. Available: https://covid19.who.int/table. [Accessed 18 January 2022].

[4] K. Yan, H. Yan and R. Gupta, "The predicted trend of COVID-19 in the United States of America under the policy of "Opening Up America Again"," Infectious Disease Modelling, vol. 6, pp. 766-781, 2021.

[5] I. Smith, "Omicron COVID variant 105% more transmissible than Delta, French scientists find," Euronews.next, 7 January 2022. [Online]. Available: https://www.euronews.com/next/2022/01/07/omicron-covid-variant-105-more-transmissible-than-delta-french-scientists-find. [Accessed 19 January 2022].

[6] N. Kohzadi, M. S. Boyd, B. Kermanshahi and I. Kaastra, "A comparison of artificial neural network and time series models for forecasting commodity prices," Neurocomputing, vol. 10, pp. 169-181, 1996.

[7] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay and et al., "Package 'forecast'," CRAN, CRAN, 2022.

[8] R. Agata and I. G. N. M. Jaya, "A comparison of extreme gradient boosting, SARIMA, exponential smoothing, and neural network models for forecasting rainfall data," Journal of Physics: Conference Series, vol. 1397, pp. 1-9, 2019.

[9] Z.-Y. Chen, T.-H. Zhang, R. Zhang, Z.-M. Zhu, J. Yang, P.-Y. Chen, C.-Q. Ou and Y. Guo, "Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China," Atmospheric Environment, vol. 202, pp. 180-189, 2019.

[10] G. E. P. Box, G. M. Jenkins, G. C. Reinsel and G. M. Ljung, Time Series Analysis Forecasting and Control Fifth Edition, Hoboken: John Wiley & Sons, Inc, 2016.

[11] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2021.