ISSN: 1999-1630

SEU Journal of Science and Engineering, Vol. 15, No. 1, June 2021

ARIMA and Exponential Smoothing Model for Forecasting the Outbreak of COVID-19 in Bangladesh

Md Mijanur Rahman[1,*] and Md Jakaria Hossain Ridoy[1]

[1]*Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh*

Abstract

The world is facing a tough situation right now because of an outbreak of a lethal virus called SARS Cov-or COVID-19. Lots of people are being infected by the virus. A great number of people are dying because of severe health conditions after being infected with the virus. This virus has caused a significant negative impact on the day-to-day activities of human life as well as in the world economy. Unfortunately, no vaccines have been invented now to diminish the virus. Although some drugs– (e.g., chloroquine, hydroxychloroquine, and azithromycin) –have been suggested to treat COVID-19 patients. However, using those medicines often does not reduce the number of cases and death. So, now is the question of when this deadly virus will come to its end. A precise forecasting model is required to respond to this inquiry. For forecasting, it is necessary to have enough historical data. The reliability of forecasting also depends on accurate data. There are several outbreak forecasting models used by researchers around the world to make informed decisions and to take the necessary precautions beforehand. Between those models, researchers have found time-series and statistical models more precise. To estimate key epidemiological steps and to predict the potential trajectory of the epidemic, this paper will concentrate on a comparative study of ARIMA models and Exponential Smoothing (Holt's Method) models to forecast monthly COVID-19 cases in Bangladesh.

Keywords: Moving Average (MA), Auto-Regressive (AR), ARIMA, COVID-19, Exponential Smoothing, Holt's Smoothing.

I. Introduction

COVID-19, a word-wide catastrophe, is an infectious virus that spreads across the world and is responsible for contaminating millions and causing death to a vast number of people of all ages after primary deaths were confirmed at the end of 2019. In about a few months, the disease spread easily around the globe, arriving at a sum of roughly 12.7 million confirmed cases and 563,296 death cases starting on 31 December (John Hopkins, 2020). On 20th March 2020, WHO (World Health Organization) stated the spread of COVID-19 as a worldwide pandemic.

According to WHO, the virus is transferred from one person to another. The infected person may start showing symptoms within 14 days, depending on the incubation period, or may not show any symptoms. Also, from WHO the clinical signs of mild to moderate cases include dry coughs, nausea, and fever, and in extreme cases, shortness of breath, fever, and tiredness can occur. Individuals with pre-existing conditions such as diabetes, heart disease, and so on are more susceptible to the virus. Currently, there are no vaccines for preventing deadly virus attacks. The drugs like hydroxychloroquine are used for symptomatic treatment (Schraer, 2020). According to WHO, people are being suggested to wash their hands with soap for at least 20 s and avoid close contact with people.

The first case of COVID-19 in Bangladesh was started in Dhaka on the 3rd of March 2020. By 11 July 2020, Bangladesh had registered 181,129 confirmed cases, and 2,305 deaths (John Hopkins, 2020). The ongoing worldwide COVID-19 pandemic has shown a nonlinear and complex nature. For this purpose, reasonable forecasting of possible reported cases is taking place and this is playing a vital role in developing a strategic plan and interventions to avoid much worse situations.

Various statistical approaches like the Time Series model (Kurbalija *et al*., 2014), multivariate

* **Corresponding Author:** Md Mijanur Rahman, Assistant Professor, Department of Computer Science and Engineering, Southeast University, House # 64, Road # 18, Banani, Dhaka 1213, Bangladesh; Email: *mijanur.rahman@seu.edu.bd*

linear regression (Thomson *et al*., 2006), gray forecast simulations (Wang *et al*., 2018a; Zhang *et al*., 2017), wavelength models (Bulut *et al*. 2020) have recently been used to forecast epidemics. But because of the randomness of epidemics above mentioned statistical methods are insufficient for predicting the trends and generalizing the outcome.

Autoregressive Integrated Moving Average (ARIMA) models are the most widely used statistical models and are considered the best at predicting infectious diseases such as malaria (Anokye *et al*. 2018), measles (Sharmin and Rayhan 2011, Ceylan, 2020), Alzahrani *et al*., 2020). ARIMA concept has been successfully extended to any kind of disease trend forecasting also in other areas because it's pretty straightforward to explain to the end-user (Cao *et al*., 2020), implementation is fast and for its ability to illustrate the data collection. The ARIMA model is widely used for forecasting purposes with the time series data. Mostly, it deals with nonstationary time series (e.g., the series data mean and standard deviation changes over time) to capture the linear pattern of an epidemic pandemic. It forecasts the value of the potential time series by recognizing the values of the prior time series of its own and the error from an earlier point of time. In our study, several ARIMA models have been formulated with different sets of ARIMA parameters (e.g., ARIMA (2, 1, 0), ARIMA (2, 2, 2), ARIMA (0, 2, 1)). ARIMA (2, 2, 2) was chosen as the best model for assessing the prevalence trends of COVID-19 in Bangladesh.

Various analytical, computational and machine learning models have been used in recent studies to predict the incidence, prevalence, and mortality rates of COVID-19 worldwide as well as in countries such as China, France, Italy, Spain, Brazil, Germany, Turkey, Saudi Arabia, and others. For example, Cylan (2020) developed an ARIMA model with a data-driven approach to estimate the prevalence of COVID-19 in Italy, Spain, and France. Alzahrani *et al*. (2020) have employed the ARIMA model to forecast the daily new cases of COVID-19 for about four weeks in Saudi Arabia. Singh *et al*. (2020) analyzes dynamic models with a data-driven approach of advanced ARIMA models to produce 20-day forecasts of accumulated reported deaths and recoveries from COVID-19 by region, territory for the top 15 affected COVID-19 countries. Ribeiro *et al*. (2020) analyzed and explored the predictive ability of machine learning regression and statistical models for one, three and six days in advance of combined events of COVID-19 cases in Brazil with models like ARIMA, Cubist regression (CUBIST), SVR (Support Vector Machine), LR (Linear Regression) (Random Forest). Roda et al. (2020) compared the conventional SIR and SEIR systems with the COVID-19 model in Wuhan Province, China (Roda *et al*., 2020). Wang *et al*. (2020) developed a Patient Information Based Algorithm for estimating the death rate of COVID-19 in real-time using publicly available data from Wang *et al*. (2020). Rustom *et al*. (2020) analyzed four forecasting models for predicting new cases, death rate showing Exponential Smoothing models as best over LR, LASSO, and SVM models. Yonar *et al*. (2020) modeled the data using some curve estimation approach and proposed Box- Jenkins, Holts, and Brown linear exponential smoothing for forecasting newly infected cases.

This paper aims to forecast the outbreak of COVID-19 cases in Bangladesh for one month as well as to identify the best model to forecast the COVID-19 pandemic among ARIMA and Exponential Smoothing (Halt's Method). Health care authorities can gain crucial information about the intensity of the epidemic at peak times from this study. They will have ideas about materials needed for patients across the country shortly.

II. Related Works

The comparison of ARIMA and GM (1,1) is used for the prediction of the hepatitis B forecast. The author found that the ARIMA model has shown the best outcome over the GM (1,1) model. The root means square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) of ARIMA (3,1,1) (0,1,2) 12 model was lower than GM (1,1) model on the fitting part and forecasting part (Wang *et al*., 2018).

Zheng *et al*. (2015) showed that the ARIMA model is an important forecasting model to predict dangerous diseases. ARCH models are a kind of similar tool used to deal with time series if the variability of the random disturbance is different

across elements of the vector. The author establishes ARIMA (1, 1, 2) (1, 1, 1) 12 model and ARIMA (1, 1, 2) (1, 1, 1)12-ARCH (1) model can be used to forecast the morbidity of TB in Xinjiang, China. Modified analyses show that the blended model has better performance.

In their article by Bulut and Yildiz (2020), the wavelength model has been used to predict the magnitude of the COVID-19 pandemic. The authors have used two different data sets for predicting the first 36 days of the pandemic. The authors used data mining techniques to combine the two data sets and Microsoft Excel and R programming to predict the pandemic situation.

A regressive Support Vector Regression (SVR) has been used to forecast the monthly outbreak of measles. The study transformed the data into some features that may reflect in the measles outbreak. This paper also shows that the efficiency of the model increases when window sizes vary. The authors showed that lower window size generates low mean error but higher window sizes generate a higher mean error and reflected the best outcome of the model (Uchenna *et al*., 2020).

The research focuses on a detailed statistical analysis of Covid-19 cases as well as on the forecast of newly infected cases in all countries. The authors modeled the data using a curve estimation approach and then used Box-Jenkins, Brown, and Holt linear exponential smoothing to estimate newly infected cases (Yonar *et al*., 2020). This research demonstrates the ability of ML models to estimate the number of future patients affected by COVID-19. The authors analyzed four forecasting models Like Linear Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Vector Machine Support (SVM), and Exponential Smoothing (ES). Each model makes three types of predictions, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The author pointed out that the Exponential Smoothing is doing well following the LR and LASSO models whereas the SVM performs very poorly on all predictions.

Abebe (2020) analyzed that Holt's Method is better than the exponential growth and single exponential smoothing to estimate potential cases of COVID-19 in Ethiopia using the Root Mean Sum of Square Error (RMSSE) to assess the effectiveness of these models. The study showed that new predicted cases had an exponential growth rate for the next three weeks in Ethiopia.

III. Methodology

Data collection (A) and analysis (B) have been performed for analyzing and validating the forecasting of the COVID-19 case.

A. Data Collection

The data has been available to the public ever since the WHO (World Health Organization) declared Covid-19 a global epidemic. So that by studying this deadly virus and its dangerous consequences, one can predict the future horrors. Many public sites are providing access to these data at no cost. We have collected a dataset from an open-source coronavirus data source *https://ourworldindata.org* (Ritchie, *et al*., 2020). The data source is being regularly updated. Those data sets consist of confirmed cases, death cases, testing.

B. Data Preprocessing

Data preprocessing is a crucial part of any data analyzing method or prediction. Nan and empty values have been filtered out from the data as data may contain Nan values or empty values or unnecessary data. Check for stationarity of the data has also been performed. There are several steps involved in analyzing data. Figure 1 shows the methodology diagram.

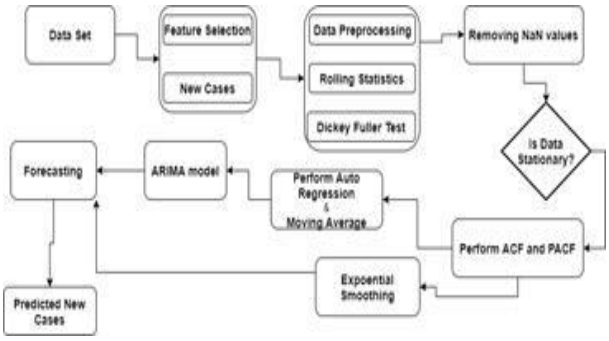

Figure 1: Methodology Diagram

C. Model Selection

Forecasting with time series needs some historical data. If there is no data or lack of data, then forecasting has been made depending on the judgmental approach, which does not always predict or forecast reliable output. Our data set can

be fit into the well-known forecasting methods ARIMA model as the model works well for historical data. Box and Jenkins first developed the model in the 1970s. It's the Auto-Regressive Optimized Moving Average Model principally a more suitable combined version of the Auto-Regressive Moving Average (ARMA) model. The AR represents Auto regression which depends on its past values. The MA uses historical projection errors in regression as a model rather than past estimation vector values. The ARIMA model represents Non- mainly a more suitable combined version of Auto Regressive Moving Average (ARMA) model. The AR represents Auto Regression which basically depends on its past values. The MA uses past forecast errors in a regression like a model rather than using past values of the prediction variable. The ARIMA model represents Non- Seasonal as ARIMA (*p*, *d*, *q*), where *p* represents the direction of the auto regressive model, *d* represents how many times subtracted from the past values, and *q* represents the moving average model. In our ARIMA model, we have used the order of ARIMA (0, 1, 2) for AR and MA, and for ARIMA we choose ARIMA (2, 1, 0). We also choose the Exponential Smoothing model to compare and contrast the forecasted output with the ARIMA model. As COVID-19 cases are increasing day by day (e.g., trend in the data and no seasonal factor) so it is clear that Holt's Linear Exponential Smoothing method can be used in our data sets.

IV. Results and Discussions

Google Colab was used for getting the output predicted by the model, which is a free cloud service provider and does support GPU processing. With this tool, one can do Machine Learning, Deep Learning, or any forecasting work. Google Colab provided us with 12 GB RAM and 100 GB SSD shareable storage when we connected our application to their server.

For more information and accuracy, the data set has been divided into one-third of the data. For the pre-part, we used the attest prediction model and matched the outcome with the final result to see how well the model is forecasting the outbreak. COVID-19 data sets for Bangladesh have been taken from *https://ourworldindata.org* (Ritchie, *et al*., 2020). Date starting from 06 April 2020 as per Fig. 2, in Bangladesh, the first COVID-19 case was recorded on 15 March 2020. When accessed the data set was recorded from 6th April 2020 to 6th October 2020.

date	new_cases
2020-04-06	18
2020-04-07	35
2020-04-08	41
2020-04-09	54
2020-04-10	112
2020-04-11	94
2020-04-12	58
2020-04-13	139
2020-04-14	182
2020-04-15	209

date	new_cases
2020-09-27	1106
2020-09-28	1275
2020-09-29	1407
2020-09-30	1488
2020-10-01	1436
2020-10-02	1508
2020-10-03	1396
2020-10-04	1182
2020-10-05	1125
2020-10-06	1442

Figure 2: Sample data from head and tail of the data set

Initially, a graph for confirmed cases as in Fig. 3 was shown where the X-axis represents the date, and the Y-axis represents the new cases. Some pre-task had to perform to get a good outbreak forecast. Among them checking for stationarity of data is important in time-series analysis. There are two ways to check stationarity in time-series data:

i. Rolling Statistics
ii. ADCF (Auto Distance Correlation Function)

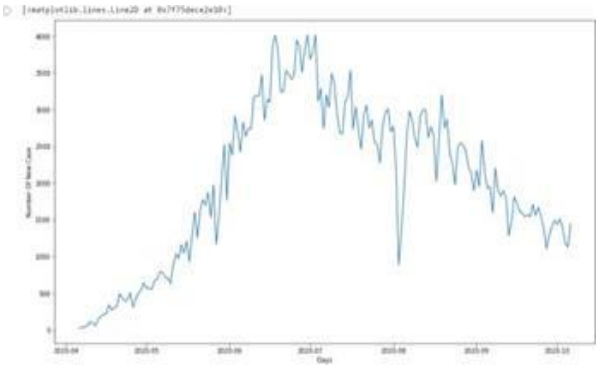

Figure 3: Plot value for row data frame

In rolling statistics, the mean and standard deviation have been found using a window size of From Fig. 4 (a), it is easily seen that data is not stationary data as the mean and standard deviation are not constant which indicates to proceed to the next step.

After this, another important test called the dickey fuller test has been performed to find the

performance of the data. The test required Ad fuller's function. In this function, we put the raw data to find these four and some critical values of different percentages. The results are shown in Table 1.

Table 1: Dickey-Fuller Test Table

Test Statistic	2.630103
p-Value	0.999080
Lags Uses	9.000000
Number of Ob	61.00000
Critical Value (1%)	-3.542413
Critical Value (5%)	-2.910236
Critical Value (10%)	-2.592745

Initially, the p-value was set to 0.05 but in the test, the result showed the *p*-value is 0.365248 which is a better estimation for the hypothesis. On the other hand, it is found that the critical value is less than the test statistic, and as our data is not stationary, so this defines that model is performing better.

After all those tests and validation, it's time to get more insight from the data. This time data was converted into a log scale to find more accurate values to predict the result of our hypothesis. In Fig. 4 (b), we can see the number of Y-axis is changed because of its log but trends remain the same.

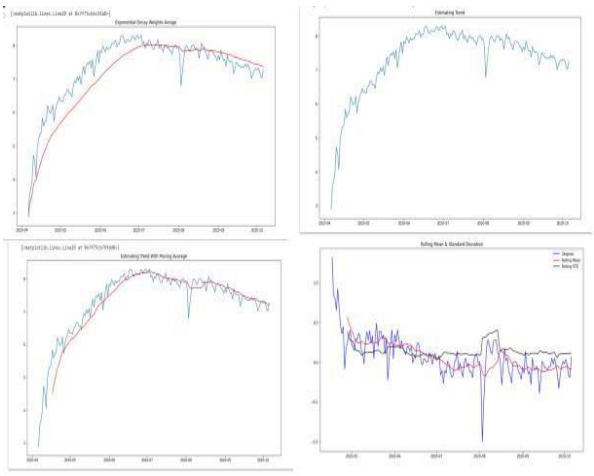

Figure 4: (a) Row Data Frame Rolling Mean and Standard Deviation. (b) Rolling Statistics in Log Scale Data Set. (c) Augmented Data Frame Rolling Mean and Standard Deviation. (d) Augmented Rolling Statistics in Log Scale Data Set.

Table 2: Augmented Dickey-Fuller Test Table

Test Statistic	-4.528792
p-Value	0.000174
Lags Uses	8.000000
Number of Ob	51.00000
Critical Value (1%)	-3.565624
Critical Value (5%)	-2.920142
Critical Value (10%)	-2.598085

All the NaN values were removed from our new log scale data set. Then an Augmented Dickey-Fuller test was performed for the log scale data set. The Adfuller's function gave an insight as shown in Figs. 4 (c)-(d). The graph shows no static value of the data set in rolling mean and standard deviation.

In the dickey fuller test, the test statistic shows a negative value and the *p*-value is near to zero which indicates our model is performing well for the data set.

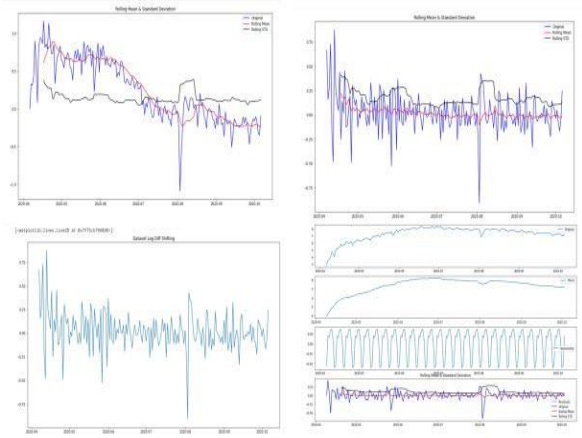

Figure 5: Exponential Decay Average

To find the weighted average the Exponential Decay Average has been used to finally find whether the dataset is stationary or not? And the result showed that the data has no stationarity.

After performing all checks, it is time to perform the actual forecasting using the time series model. In time series, seasonal decompose function, Trend, Seasonality, Residuals have been used.

In the second stage, the ARIMA model has been applied to find the predicted value. Before proceeding with the ARIMA model it is necessary to find the P and Q values. To find the P and Q

value ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) have been used with OLS (Ordinary Least Square) method. As shown in Fig. 6.

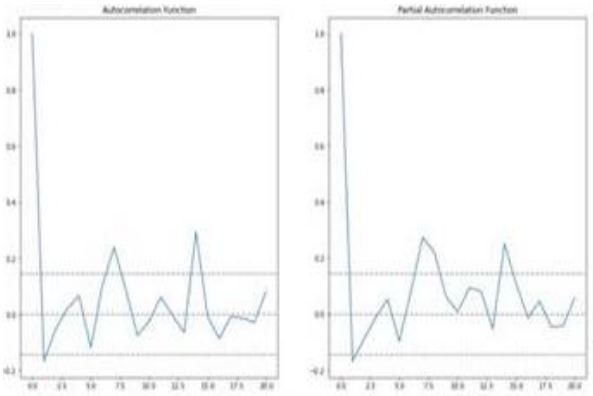

Figure 6: ACF (Auto-correlation Function) and PACF (Partial Auto-Correlation Function)

So, from the graph, it is found that the *p*-value is equal to 1, and the Q value is equal to 2. Now time to introduce the ARIMA model, but first, it is good to have RSS (Residual Somu's Square) value for the AR and MA model and finally for the ARIMA model. Placing P and Q in the AR model gave an RSS value of 7.4576 as in Fig. 7 (a) and the MA model gave an RSS value of 7.6576 as in Fig. 7 (b) later than those AR and MA values had been used in the ARIMA model to get RSS value of 7.0513.

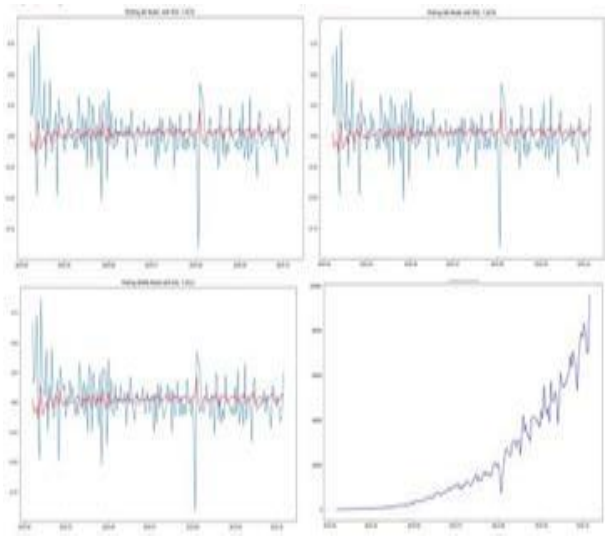

Figure 7: (a) AR (Auto-regressive) Model (b) MA (Moving Average) model (c) ARIMA Model (d) Prediction ARIMA Experimental Part

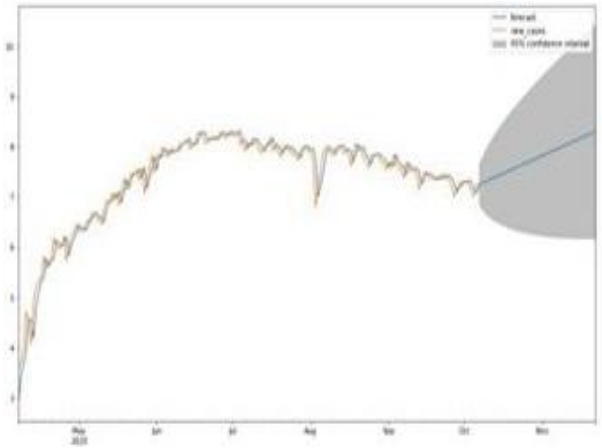

Figure 8: Final Prediction Plot

After performing all validation testing a final result of prediction is shown in the graph as in Fig. 8, where the orange line shows the new case. The blue line is the forecast value and the model accuracy or confidence interval of 95%. Finally, when a full data set is fitted into the module, then it shows that the data is not stationary and the ARIMA model RSS value is 14.2236. The model accuracy was approximately 95%.

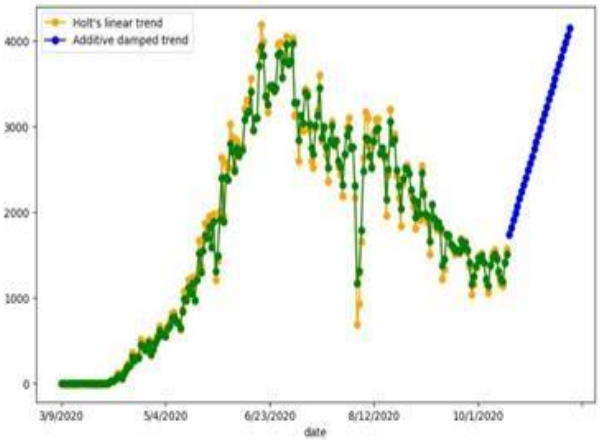

Figure 9: Exponential Smoothing

Later, after Exponential Smoothing as in Fig. 9, particularly Halt's Method was selected to forecast COVID-19 new cases. Halt's Method is good when the data set has a clear trend view. For this, we use the stat model's python module that has tons of built-in functions for statistical models. Long-term prediction with Exponential Smoothing is not very effective as forecasting will increase or decrease indefinitely in the future. The Damped Trend model can be used to restrict the indefinite behavior. In our forecast case,

1. Alpha: The level smoothing element.
2. Beta: Smoothing component of the trend.
3. Trend Type: additive or multiplier.
4. Dampen Type: additive or multiplier.
5. Phi: Coefficient of steaming.

Values have been formulated to forecast with Holt's Method (e.g., alpha - 0.8, beta- 0.3,) or (0.7, 0.2) but found (0.8 and 0.3) best for the Holt's Linear Trend. In additive, the damped trend (0.8, 0.2) produces a good result.

The damped pattern method is a method that incorporates a dampening parameter such that the trend can converge to a fixed value in the future (it flattens the trend). The forecast shows that the damped trend has a high constant trend of forecast in the future where the ARIMA model has a clear lower linear trend of forecast and 95% confidence, so it is certain that ARIMA can be a good choice of forecasting for the data set.

V. Conclusions

The paper did not consider the vaccination process and any way to prevent the pandemic. This study mainly focused on the monthly outbreak forecasting for the local government. The ARIMA and Exponential Smoothing models were used for predicting the monthly new confirmed cases. The output of the study shows that the ARIMA model performs best on the given data sets for monthly COVID-19 prediction. Based on the findings from the study of these two models, the newly infected number of people will increase in the coming month. This study will help the government or people directly involved in coronavirus treatment to take the necessary steps to follow beforehand. The efficiency can be improved by providing a lot more data. And this study can be further improved by categorizing whether a male will be more affected or females as well as improving the ARIMA parameters and smoothing factors for exponential smoothing. The model can also be used in other infectious diseases outbreak predictions.

References

"COVID-19 Map," Johns Hopkins Coronavirus Resource Center, 2020. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed: 19 July 2020].

Brockwell PJ, Davis RA, "Introduction to Time Series and Forecasting," 2001, Second Edition, Springer.

C. Bulut and Y. Kato, "Epidemiology of COVID-19," Turkish Journal of Medical Sciences, vol. 50, no. -1, pp. 563-570, 2020. Available: 10.3906/sag-2004-172 [Accessed 20 October 2020].

H. Ritchie, "Corona virus Source Data," Our World in Data, 2020. [Online]. Available: https://ourworldindata.org/coronavirus-source-data?fbclid=IwAR2pt5Vlfz9hcX92bnPEGc4NyFvGlCvocuVfAH-kH6ElRefHWhd73UnTCTU. [Accessed: 19 July 2020].

H. Yonar, A. Yonar, M. Tekindal and M. Tekindal3, "Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods," Eurasian Journal of Medicine and Oncology, 2020. Available: https://www.ejmo.org/10.14744/ejmo.2020.28273/. [Accessed 20 October 2020].

M. Ribeiro, R. da Silva, V. Mariani and L. Coelho, "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil," Chaos, Solitons and Fractals, vol. 135, p. 109853, 2020. Available: 10.1016/j.chaos.2020.109853 [Accessed 20 October 2020].

N. Uchenna, A. Oreoluwa, O. Rotimi, B. Oluwatobi, and A. James, "Forecasting Infectious Disease Outbreak Using Support Vector Regression (SVR) Case Study: Measles (Rubeola)," PhD, Babcock University, Federal Polytechnic Ilaro, 2020.

Q. Li *et al*., "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia," New England Journal of Medicine, vol. 382, no. 13, pp. 1199-1207, 2020. Available: 10.1056/nejmoa2001316 [Accessed 20 October 2020].

Q. Li, W. Feng and Y. Quan, "Trend and forecasting of the COVID-19 outbreak in China," Journal of Infection, vol. 80, no. 4, pp. 469-496, 2020. Available: 10.1016/j.jinf.2020.02.014 [Accessed 20 October 2020].

R. Anokye, E. Acheampong, and I. Owusu, "Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence," PhD, Edith Cowan University, Kwame Nkrumah University of Science and Technology, 2018.

R. Schraer, "Coronavirus: Hydroxychloroquine trial begins in the UK," BBC News, 2020.

S. Alzahrani, I. Aljamaan and E. Al-Fakih, "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions," Journal of

Infection and Public Health, vol. 13, no. 7, pp. 914-919, 2020. Available: 10.1016/j.jiph.2020.06.001 [Accessed 20 October 2020].

Sharmin S. Rayhan I., "Modeling of infectious diseases for providing a signal of epidemics: A measles case study in Bangladesh," Journal of Health, Population, and Nutrition, vol. 29, no. 6, 2011, pp. 567-573, https://doi.org/10.3329/jhpn.v29i6.9893.

T. Abebe, "Forecasting the Number of Coronavirus (COVID-19) Cases in Ethiopia Using Exponential Smoothing Times Series Model," medrxiv, 2020. Available: https://www.medrxiv.org/content/10.1101/2020.06. 29.20142489v1. [Accessed 20 October 2020].

T. Bulut and M. Ç. Yildiz, "Prediction of Size of the COVID-19 Pandemic Using Wavelength Models: Cases of Turkey and World," 2020.

T. Tajmim, "Bangladesh recommends controversial drugs for Covid-19 treatment," TBS News, p. 1, 2020.

V. Kurbalija *et al*., "Time-series analysis in the medical domain: A study of Tacrolimus administration and influence on kidney graft function," Computers in Biology and Medicine, vol. 50, pp. 19-31, 2014. Available: 10.1016/j.compbiomed.2014.04.007 [Accessed 20 October 2020].

Wang Y, Shen Z, Jiang Y, "Comparison of ARIMA and GM (1,1) models for the prediction of hepatitis Bin China," PLoS ONE, vol. 13, no. 9, pp. 1-11, 2018, https://doi.org/10.1371/journal.pone.0201987.

Z. Ceylan, "Estimation of COVID-19 prevalence in Italy, Spain, and France," Science of The Total Environment, vol. 729, p. 138817, 2020. Available: https://www.sciencedirect.com/science/article/pii/S0048969720323342#bb0150. [Accessed on 20 October 2020].

Zheng YL, Zhang LP, Zhang XL, Wang K, Zheng YJ, "Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China," PLoS ONE, vol. 10, no. 3, 2015, pp. 1-3. https://doi.org/10.1371/journal.pone.0116832.