Use the ARIMA and EMG Models to Predict the Confirmed Cases and ICU Bed Occupancy Data of COVID-19 in Massachusetts and Compare the Accuracy of the Two Models

Xinrui Cheng

Department of Boston University, MA, 02134, USA

xcheng22@bu.edu

Abstract. Beginning in 2019, COVID-19 has invaded our lives and continues until now. People are still being effort made to control the spread of the virus and treat patients. Two variables in COVID-19 data are confirmed data and ICU bed occupancy data. Collect the data, use the ARIMA model and EMG model to predict the data, and then compare it with the actual data. Through comparison, it is found that there is a small gap between the ARIMA predicted value and the actual data, which can be used to predict the data of these two variables. In the process of fighting against COVID-19, it is very important to strengthen medical service capabilities, adequate treatment resources, and strengthen the construction of ICU wards. The construction of ICU is very complicated and involves many medical equipment, including ventilator, infusion pump, electrocardiogram machine, extracorporeal membrane oxygenation machine (ECMO), oximeter and so on. Therefore, the accuracy of data prediction is very important. Sufficient ICU beds and medical resources can be prepared according to the forecast. Do a good job of safeguarding the treatment work and medical services to ensure the level and quality of patient treatment and avoid causing panic.

Keywords: Covid-19, ARIMA model, EMG model, Confirmed data, ICU.

1. Introduction

The ongoing COVID-19 pandemic has been lasting for years and infected millions of victims since December 2019. Due to its high infectious rates and severe syndromes (e.g. deaths, respiratory problems, and weak immune system) and lack of effective vaccines and medicine, it is very important to precisely forecast the spreading pattern of the epidemic to help the government authorities to formulate appropriate public epidemic prevention policies (e.g. containment, lockdown, and mask) at the lowest social cost for the entire society. For example, if the infectious rate for the future month is quite high, then more extreme containment policies should be planned. On the contrary, if the symptoms are minor and death rates are quite low, less strict containment policies should be proposed.

In terms of research methods, the paper will first conduct a descriptive analysis of the collected COVID-19 epidemic data such as visualization of summary information like means, quantiles, variance, and cross-tabulations. Then the time series data will be analyzed by the following forecast models: ARIMA, and EMG model. The potential trend and seasonality features of the time series data will be investigated in detail. The effectiveness and exactness of these models will be evaluated through various metrics. It is expected that some epidemiological information will be discovered, and the forecast models will be directly helpful for the government and public health authorities to select appropriate containment policies and plans according to the specific situation of the disease.

The error between the data predicted by ARIMA model and the actual data is very small, so it is suitable for predicting the number of confirmed cases and ICU beds in COVID-19. However, EMG model is difficult to find peak value, so it is not applicable to predict the number of confirmed cases and ICU beds in COVID-19.

2. Literature review

The exponential modified Gaussian (EMG) can provide better fitting for the changes of intermittent time and protein expression [1]. ARIMA and Prophet time series prediction models are

used to predict the world and 10 countries most affected by COVID-19: the United States, Spain, Italy, France, Germany, Russia, Iran, the United Kingdom, Türkiye and India [2]. Using a simplified mathematical method, forecasting the demand for COVID-19 ICU beds, formulated emergency plans, and made resource allocation [3]. Use ARIMA model to get COVID-19 data of Hubei, China, and then predict the confirmed cases of COVID-19 in Italy in the next ten days [4]. Using Statistical and Machine Learning Models predict COVID-19 data [5]. Compared with the accuracy of ARIMA, ES and GRNN models through the daily data of new COVID-19 cases in the United States and India [6]. Prediction of COVID-19 cases based on social distance in Maryland, USA using SDI model [7]. Using ARIMA model to predict the incidence rate and mortality of hand foot and mouth disease in Sichuan Province, China [8]. Using ARIMA model to predict the temporal dynamics of COVID-19 in global transmission [9]. Using ARIMA model to predict the data of confirmed cases, death and recovery of COVID-19 in the top 15 Affected Countries [10]. Using ARIMA model to predict the data of COVID-19 in three countries: Italy, Spain, and France [11].

3. Exploratory Data Analysis

3.1. Variable

The variable is the number of confirmed cases of COVID-19 every day, and the number of ICUs occupied by COVID-19.

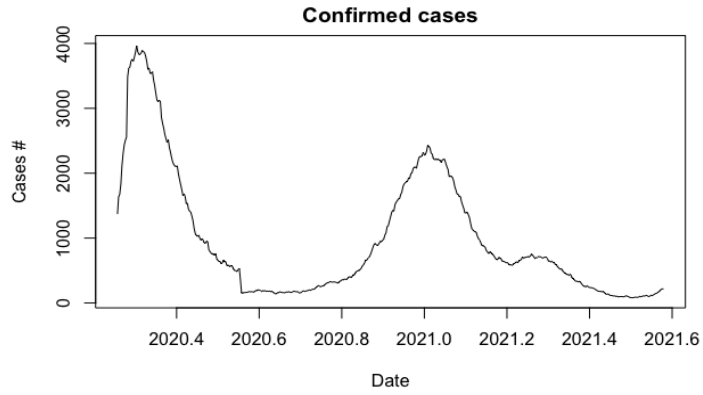


Figure 1. Confirmed cases (Photo credit: Original)

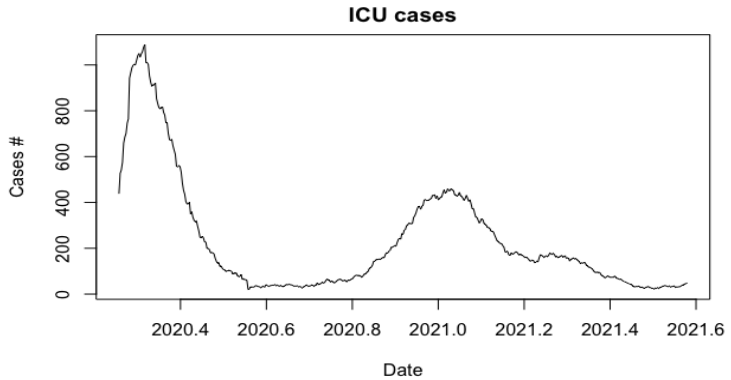


Figure 2. ICU cases (Photo credit: Original; Source: Massachusetts Government [15])

Both data present changing variance, thus it is may be necessary to consider Box-Cox transformation.

Table 1. Box-Ljung Test

Box-Ljung Test	X-squared	Degree of freedom	P-value
Confirmed	10935	50	0.0000
ICU	10619	50	0.0000

Table 2. Aufmented Dickey-Fuller Test

Aufmented Dickey-Fuller Test	Dickey-Fuller	Lag order	P-value
Confirmed	-3.1101	7	0.1048
ICU	-3.2542	7	0.07863

The ADF tests show that there are no significant evidences for unit root, i.e. both data are non-stationary data. Therefore, the ARIMA model should be a good candidate model.

The Ljung-Box tests demonstrate that there are significant evidence of non-zero serial correlation, i.e. both data are not white noise. Thus, the further modeling and analysis is meaningful.

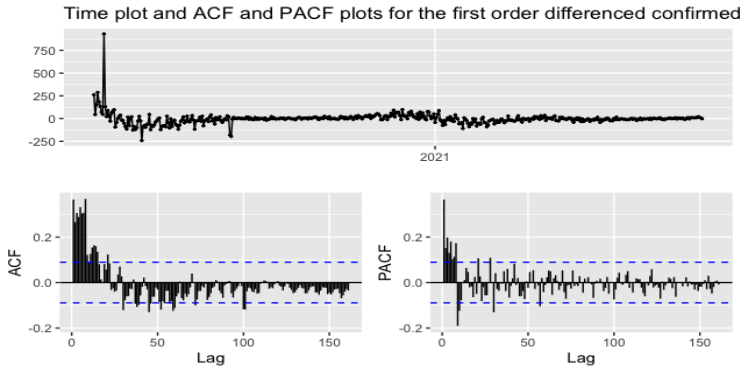


Figure 3. Time plot and ACF and PACF plots for the first order different confirmed cases (Photo credit: Original)

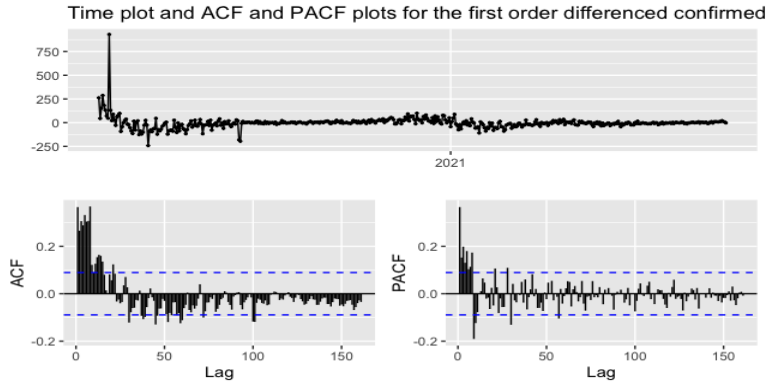


Figure 4. Time plot and ACF and PACF plots for the first order different ICU cases (Photo credit: Original)

The first order difference of both data shown above are generally stationary, although there seem to be more extreme spikes on the ICU data ACF plot.

3.2. Model

The forecasts in this paper is based on two different models, EMG and ARIMA.

3.2.1 EMG

The exponentially modified Gaussian (EMG) is capable of modeling time-series data with skewed peaks [1]. The R package IPPD provides functions for EMG modeling on intensities data [14].

$$G(t) = \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left(-\frac{(S-t)^2}{2\sigma^2}\right) \tag{1}$$

$$E(t) = k \cdot \exp(-k \cdot t) \tag{2}$$

$$n(t) = G(t) * E(t) = \frac{k}{\sigma\sqrt{2\pi}} \int_0^t \exp\left(-\frac{(t-S-\acute{t})^2}{2\sigma^2}\right)\exp(-k\acute{t})d\acute{t}$$

$$= \frac{k}{\sigma\sqrt{2\pi}} \int_0^t \exp\left(-\frac{(\acute{t}-S)^2}{2\sigma^2}\right)\exp[-k(t-\acute{t})]d\acute{t} \tag{3}$$

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,dt \quad \text{Where} \quad z = \frac{S-t}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\frac{1}{k}} \tag{4}$$

$$n(t) = \frac{\sqrt{2\pi}}{2\frac{1}{k}}\exp\left(\frac{S-t}{\frac{1}{k}} + \frac{\sigma^2}{2(\frac{1}{k})^2}\right)\left[\frac{k}{|k|} - \text{erf}\left(\frac{S-t}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\frac{1}{k}}\right)\right] \tag{5}$$

Where G(t): gaussian probability density
E(t): exponential probability density
n(t): total intermitotic time [1]

3.2.2 ARIMA

ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends [12].

$$(1 - \sum_{i=1}^{p} \emptyset_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \tag{6}$$

Used ARIMA model with the order (3, 1, 2) to forecast the COVID-19 confirmed and order (4, 1, 3) to forecast ICU cases. Then confirm whether the model is suitable for prediction by comparing the real resin and predicted value.

3.2.3 ARIMA model on confirmed cases data

Following figure presents the accuracy results of ARIMA model (p=3, d=1, q=2) on the confirmed cases data after Box Cox transformation. It is found that the model performs much worse on testing data than training data. For example, the RMSE on test set is about 20 times the RMSE on training set, and the MAE on test set is about 31 times the MAE on training set. Thus, it is concluded that the model has a poor accuracy on the confirmed data.

Table 3. Training set and test set for confirmed cases

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's u
Training set	-2.651182	59.75468	29.5483	-0.0993782	3.533829	NaN	-0.02104187	NA
Test set	-1050.862875	1200.37612	11050.8629	-755.1393574	755.139357	NaN	0.98051078	203.9938

The model ARIMA: p=3, d=1, q=2 with drift seems to fit well on the train data. The Box Cox transformation with lambda = 1 is done to stabilize the variance. There are three components: (1) autoregressive terms with order 3, (2) nonseasonal differences of order 1 for stationarity, and (3) 2 lagged forecast errors. The RMSE 59.75 on training data is not very well but acceptable.

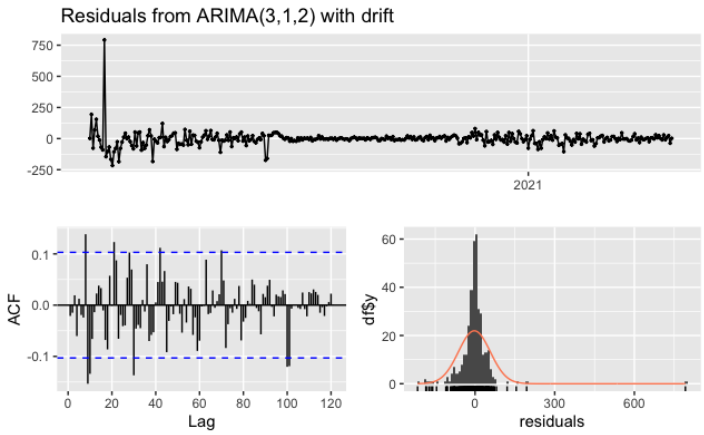


Figure 5. Residuals fro ARIMA (3, 1, 2) with drift for confirmed (Photo credit: Original)

There seem to be extreme spike values out of the error limits in the residuals. The histogram plot shows that the residuals are a little skewed from normally distribution. Thus, it is necessary to do a box-cox transformation on the data.

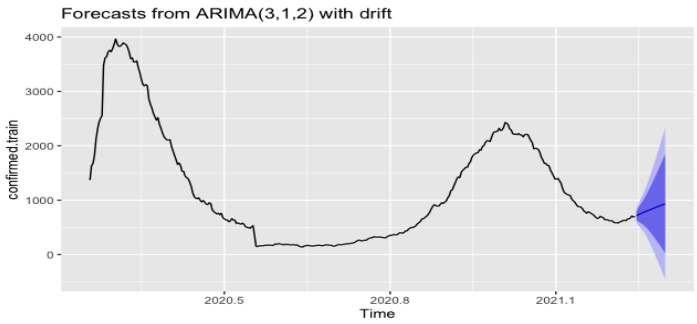


Figure 6. Forecasts from ARIMA (3, 1, 2) for confirmed (Photo credit: Original)

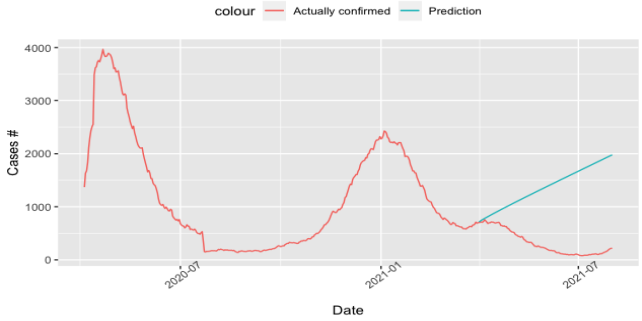


Figure 7. ARIMA (3, 1, 2) confirmed actually compare with prediction (Photo credit: Original)

The forecast values for test data present an upward trend. However, the trend of actual confirmed test data have an opposite direction. The model has a bad predicting performance! The mean value of prediction of confirmed cases is 1364, and the RMSE of prediction of confirmed cases is 1197, indicating that the model's predicting accuracy is poor.

3.2.4 ARIMA model on ICU cases data

Following figure presents the accuracy results of ARIMA model (p=4, d=1, q=3) on the ICU cases data after Box Cox transformation. It is also found that the model performs much worse on testing data than training data. For example, the RMSE on test set is about 10 times the RMSE on training set, and the MAE on test set is about 15 times the MAE on training set. Thus, it is concluded that the model has a poor accuracy on the ICU data.

Table 4. Training set and test set for confirmed cases

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's u
Training set	-0.5189501	14.83709	8.835908	-0.01252059	5.300549	NaN	-0.02544708	NA
Test set	-137.5724917	157.03145	137.593008	-333.68223935	333.693709	NaN	0.9831908	64.90347

The model ARIMA: p=4, d=1, q=3 without drift seems to fit well on the train data. The Box Cox transformation with lambda = 1 is done to stabilize the variance. There are three components: (1) autoregressive terms with order 4, (2) nonseasonal differences of order 1 for stationarity, and (3) 3 lagged forecast errors. The RMSE 14.83 on training data indicates good accuracy on training data.

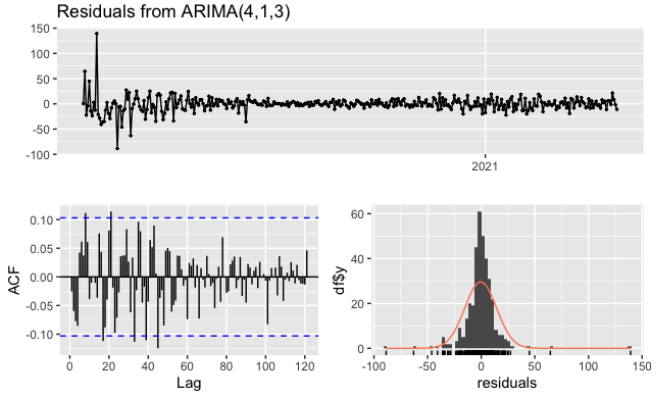


Figure 8. Residuals from ARIMA (4, 1, 3) for ICU (Photo credit: Original)

There seem to be extreme spike values out of the error limits in the residuals. The histogram plot shows that the residuals are a little skewed from normally distribution. Thus, it is necessary to do a box-cox transformation on the data.

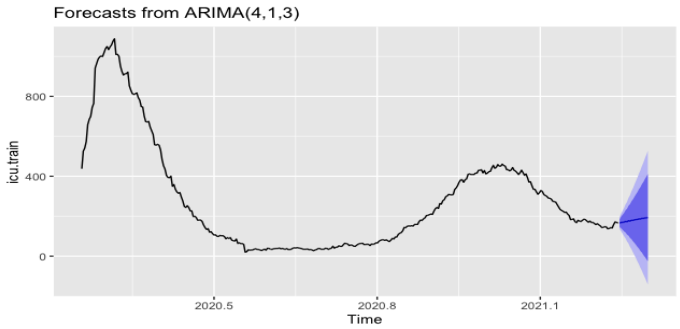


Figure 9. Forecasts from ARIMA (4, 1, 3) for ICU (Photo credit: Original)

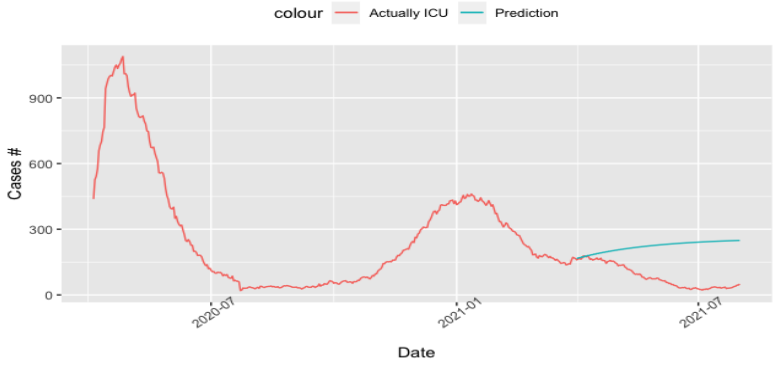


Figure 10. ARIMA (4, 1, 3) for ICU actually compare with prediction (Photo credit: Original)

The forecast values for test data present an upward trend. However, the trend of actual ICU test data have an opposite direction. The model has a poor predicting performance! The mean value of prediction of confirmed cases is 221, and the RMSE of prediction of confirmed cases is 157, indicating that the model's predicting accuracy is quite poor.

4. Results

Through ARIMA model prediction, it is found that the difference between the predicted data value and the actual data value is very small. ARIMA model is suitable for predicting confirmed cases and ICU cases.

Following figure plots the training data of confirmed cases and ICU cases (2020.4-2021-8). It is obvious that they present similar varying trends, which means EMG model could be a candidate model to capture the peak patterns [13].

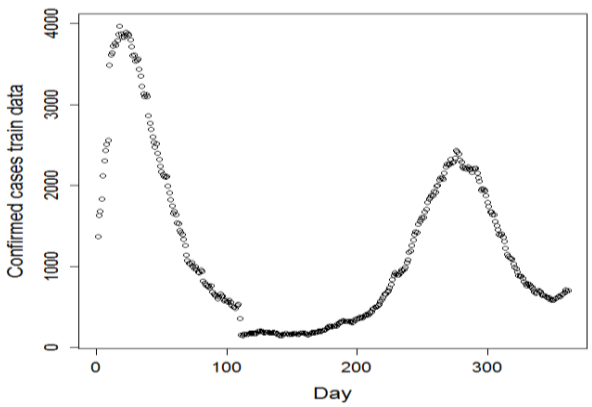


Figure 11. Confirmed cases train data by day (Photo credit: Original)

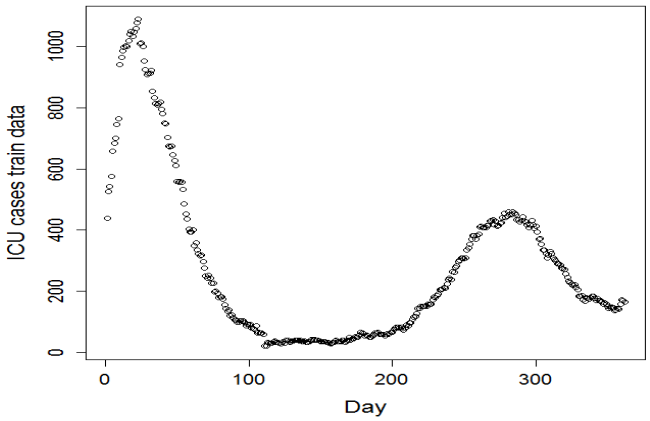


Figure 12. ICU cases train data by day (Photo credit: Original)

5. Conclusion

By comparing the predicted value with the actual value, the error of the ARIMA model is very small, and it can be used to predict the number of confirmed cases of the COVID-19 and ICU bed occupancy data. The prediction data error is small, and the ICU bed is well prepared, which can guarantee the level and quality of patient care to the greatest extent.

References

[1] Golubev. (2010). Exponentially modified Gaussian (EMG) relevance to distributions related to cell proliferation and differentiation. Journal of Theoretical Biology, 262(2), 257–266.

[2] Kumar, & Susan, S. (2020). COVID-19 Pandemic Prediction using Time Series Forecasting Models.

[3] Manca, Caldiroli, D., & Storti, E. (2020). A simplified math approach to predict ICU beds and mortality rate for hospital emergency planning under Covid-19 pandemic. Computers & Chemical Engineering, 140, 106945–106945.

[4] Yang, Wang, J., Ma, H., & Wang, X. (2020). Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy. Journal of Infection and Public Health, 13(10), 1415–1418.

[5] Chowdhury, Rafiqul I., M. Tariqul Hasan, and Gary Sneddon. "Regressive Class Modelling for Predicting Trajectories of COVID-19 Fatalities Using Statistical and Machine Learning Models." Bulletin of the Malaysian Mathematical Sciences Society (2022): 1-16.

[6] Wang, Wu, T., Wei, W., Jiang, J., An, S., Liang, B., Ye, L., & Liang, H. (2021). Comparison of ARIMA, ES, GRNN and ARIMA–GRNN hybrid models to forecast the second wave of COVID-19 in India and the United States. Epidemiology and Infection, 149.

[7] Cruz-Cano, Ma, T., Yu, Y., Lee, M., & Liu, H. (2022). Forecasting COVID-19 Cases Based on Social Distancing in Maryland, USA: A Time–Series Approach. Disaster Medicine and Public Health Preparedness, 16(5), 1798–1801.

[8] Liu, Luan, R. S., Yin, F., Zhu, X. P., & Lü, Q. (2016). Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model - CORRIGENDUM. Epidemiology and Infection, 144(1), 152–152.

[9] Chyon, Suman, M. N. H., Fahim, M. R. I., & Ahmmed, M. S. (2022). Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. Journal of Virological Methods, 301, 114433–114433.

[10] Singh, Rani, M., Bhagavathula, A. S., Sah, R., Rodriguez-Morales, A. J., Kalita, H., Nanda, C., Sharma, S., Sharma, Y. D., Rabaan, A. A., Rahmani, J., & Kumar, P. (2020). Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model. JMIR Public Health and Surveillance, 6(2), e19115–e19115.

[11] Ceylan. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. The Science of the Total Environment, 729, 138817–138817.

[12] https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp.

[13] K.U, N., & S.D., K. (2022). Exponential-Gaussian Distribution and Associated Time Series Models: Accepted: February 2022. REVSTAT-Statistical Journal. Retrieved from.

[14] Slawski M, Hussong R, Hildebrandt A and Hein M (2014). IPPD: Isotopic peak pattern deconvolution for Protein Mass Spectrometry by template matching. R package version 1.20.0.

[15] https://www.mass.gov/info-details/archive-of-covid-19-cases-2020-2021#august-2020.