International Journal of Trends in Mathematics Education Research
Vol. 5, No. 3 (2022), pp. 333-339
ISSN 2621-8488 (online)
DOI: https://doi.org/10.33122/ijtmer.v5i3.157

SAINTIS publishing

Research Article

Prediction of the number of daily active COVID-19 in Indonesia

Hedi[1]✉, Anie Lusiani[2], Anny Suryani[3], Agus Binarto[4]

[1] Department of Energy Conversion Engineering, Politeknik Negeri Bandung, West Java, Indonesia, 40559
[2] Department of Mechanical Engineering, Politeknik Negeri Bandung, West Java, Indonesia, 40559
[3] Department of Accounting, Politeknik Negeri Bandung, West Java, Indonesia, 40559
[4] Department of Electrical Electronic Engineering, Politeknik Negeri Bandung, West Java, Indonesia, 40559
✉Corresponding Author: hedi@polban.ac.id | Phone Number: +62852-2004-3920

Received: 14 June 2022 *Revised: 24 August 2022* *Accepted: 20 September 2022* *Available online: 30 September 2022*

ABSTRACT

In Indonesia, the coronavirus disease (COVID-19) decreased from April to May 2022 and in-creased slowly from May to June 2022. Statistical predictions are needed to monitor the increase in cases of this pandemic spike, as happened at the end of February 2022. This study aims to predict the rise in the number of active COVID-19 cases by applying the autoregressive integrated moving average (ARIMA) mathematical model. and multiple linear regression (MLR). Daily observation data of active cases, new cases, recovered cases, and deaths were recorded from January to June 2022 totalling 152 observations. Then ARIMA modelling for active cases and MLR modelling for daily active case observation data that depended on new cases were carried out , recovered, and died. Furthermore, the prediction results from the two models were determined the root mean squared error (RMSE), the mean absolute error (MAE), and the mean absolute percent error (MAPE). From the calculation results, the ARIMA model is smaller than the MLR. However, the prediction of the next thirty days in the MLR model is close to the actual value, while in the ARI-MA model it is below the actual value.

Keywords: COVID-19; Prediction; ARIMA; MLR;

1. INTRODUCTION

COVID-19 in Indonesia as of June 11, 2022, reached 6,059,937 confirmed cases and 156,641 deaths. From early January to mid-January 2022, the number of active cases, new cases, and deaths decreased very significantly, but starting from mid-January to February, it increased again and the peak occurred at the end of February 2022. Furthermore, there was a significant decrease until the end of May and again experienced a slow increase until June 2022 (Indonesia COVID-Coronavirus Statistics, 2022). Some of the causes of this rapid pandemic are, la ack of transmission of information in the early stages about the causes of COVID-19, symptoms, and behaviors of the spread, people affected by pre-existing diseases, lack immunity among humans, overcrowding, and lack of adequate health facilities (Gherghel & Bulai, 2020). The most important parameter in avoiding this pandemic is social distancing (WHO.c, 2020). Another cause is the lack of mobility applied, causing an increase in the rate of spread (Bustaman, 2021).

The increase in the number of new cases depends on the number of crowds, the higher the number of crowds, the higher the rate of spread. The increase in the number of new cases in chains has the potential to increase the spread of new cases again. Without strict control management, the number of new cases will increase faster (Yonar, 2020). The role of the Indonesian government in controlling this pandemic is quite successful through the implementation of community activity restrictions (PPKM) and the provision of vaccines throughout Indonesia. Information on the forecast for an increase or decrease in this pandemic is very much needed. Statistical predictions can help in estimating the number of COVID-19 cases in the future. Various mathematical models are used to monitor the upcoming increase in the number of cases. In Modelling, the number of COVID-19 cases will be very helpful in monitoring the development of this pandemic in the short and long term. In modelling the COVID-19 case, researchers used more autoregressive integrated moving average (ARIMA) forecasting models (Zeynep Ceylan, 2020), (Ganiny & Nisar, 2021), (Kamboj et al., 2020).

In ARIMA modeling, it only involves the current data as a response variable that relies on past data as a predictor variable, so that if the data on active daily cases of COVID-19 is modeled, then this model cannot explain the influence of new cases, recoveries, and deaths. Predictive modeling involving many free variables that affect response variables is modeled with MLR (Bull, 1998), (Cannon & Mckendry, 1999). MLR modeling in COVID-19 cases is applied to express the variable relationship of the number of daily cases confirmed, recovered, new cases, and deaths. Chaurasia et al. apply a large number of deaths as a function of a large number of confirmed, recovered cases and the rate of increase (Chaurasia & Pal, 2020).

The MLR model can be used to predict air pollution problems in Kuala Terengganu, Malaysia (Abdullah et al., 2017) and Saya City, Turkey (Z Ceylan & Bulkan, 2018). In the optimization of more sustainable and renewable energy resources, MLR modeling is also used (Ali et al., 2020). Short-term load forecasting is crucial in power system operation and control. The periodic low-frequency components are predicted by the MLR method (Li et al., 2020). Studies relating to load demand forecasting at long-term peak electricity and ionospheric TEC forecasting can use multiple linear regression (MLR) methods (Al-Hamad & Qamber, 2019; Inyurt et al., 2020).

In this study, two models will be applied to predict the number of active cases of COVID-19 in Indonesia for the period January 2022 to the period of June 2022 using the ARIMA model. Furthermore, modeling the number of active cases as a response variable that depends on the number of new cases, recoveries, and deaths, as a predictor variable using the MLR model. From these two models, predictions of active cases of COVID-19 are determined in the following month. Previously, we have implemented a predictive model for the number of active cases of COVID-19 in Indonesia using ARIMA and SARIMA (Suryani & Binarto, 2021).

2. RESEARCH METHOD

This study uses a secondary data sample for the period 11 January 2022 to 11 June 2022, which was recorded by (Indonesia COVID - Coronavirus Statistics, 2022). What is analyzed in this study is modeling the number of active cases of COVID-19 using ARIMA and modeling active cases that depend on new cases, recovering, and dying using MLR.

2.1 ARIMA

COVID-19 cases have a very fast transmission nature, the number of active cases will affect the number of new cases, so strict supervision is needed in this pandemic (Zhang, et al., 2020). Many researchers apply time series data models to predict pandemic cases (Cambodia et al., 2020), (Davis et al., 2019). Research (Yonar, 2020) applies this model to predict the number of confirmed cases of COVID-19. Most researchers apply the ARIMA time series data model to predict cases of COVID-19 data (Benvenuto et al., 2020). In observing the time series data for the value of the data $y_1, y_2, y_3, \ldots y_T$, with the current observation y_T and the value of the past observations $y_{T-1}, y_{T-2}, y_{T-3}, \ldots y_1$, the equation model of the observation function y_t is determined which depends on the p - lag observations.

$$y_t = \text{f}(\ y_{t-1}, y_{t-2},\ y_{t-3}, \ldots y_{t-p}\ ) \tag{1}$$

and the error function which depends on the q - lag observations

$$y_t = \text{f}(\ \varepsilon_{t-1}, \varepsilon_{t-2},\ \varepsilon_{t-3}, \ldots \epsilon_{t-q}\ ) \tag{2}$$

If the time series data y_t with observation time t = 1, 2, 3, …T for T the number of observations. The time series y_t is stationary after going through the transformation with a difference of d times, then the combined time series data from the functions of equations 1 and 2 are modeled with ARIMA(p, d, q). with equation

$$y_t = \mu + u_t \tag{3}$$

$$\phi_p(B)(1-B)^d u_t = \theta_q(B)\varepsilon_t \tag{4}$$

Where, $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$; $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$,
μ = mean; and ε_t= residual and $B^p(y_t) = y_{t-p}$.

2.2 Identification and Estimation ARIMA Model

Time series data on the number of active cases measuring 152 were plotted to see their stationarity graphically, then statistically performed by applying the Augmented Dickey-Fuller (ADF) test. Then the logarithm transformation stabilizes the variance. If it is not stationary in the mean then *differencing*. The estimation of the *p parameter is determined* through the ACF plot pattern in the form of a sine wave or an exponential after the p lag while the PACF is in the form of the damped sine wave or an exponent at the q lag. In addition, the possible values of p and q are selected, and from these possibilities are calculated *Akaike's Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC), and the smallest possible value of p and q was chosen. A diagnostic test is done through a parameter significance test, the t distribution is used, to test the hypothesis that the ARIMA model equation coefficient (p, d, q) is equal to zero.

2.3 MLR

Based on research (Rath et al., 2020), MLR can be used to predict the number of active cases as a function of the number of positive confirmed cases, recovered, and died. In this study, modeling will be determined that connects three predictor variables with one response variable, with the y response function depending on the three predictor variables, namely

$$\text{Y} = \text{f}(X_1, X_2, X_3) \tag{5}$$

if the response is Y with T observation values , $y_1, y_2, y_3, \ldots y_T$ and predictor X_1, X_2, X_3, each of which has T observation values $x_{11}, x_{21}, x_{31}, \ldots x_{T1}$; $x_{12}, x_{22}, x_{32}, \ldots x_{T2}$; $x_{13}, x_{23}, x_{33}, \ldots x_{T2}$, then the function of equation 5 is expressed.

$$\text{Y} = \beta_o + \beta_1 \text{X}_1 + \beta_2 \text{X}_2 + \beta_3 \text{X}_3 + \varepsilon \tag{6}$$

In this study, response y stated the number of daily active cases influenced by the independent variables X_1 (daily number of new cases), X_2 (daily number of recoveries), and X_3 daily number of deaths. In matrix is expressed in the equation

$$Y = X\beta + E \quad (7)$$

By applying *ordinary least squares* (OLS), the parameter estimation

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad (8)$$

Furthermore, the diagnostic test through the parameter significance test used t distribution and analysis of variance to see if the model parameters were significant.

2.4 Model Evaluation

In this section, the best model is determined from the two models with the smallest criteria from the calculation of MAE, RMSE, and MAPE. Next, plot the predictions of the results of the two models which are compared with the actual data the following month

3. RESULTS AND DISCUSSION

The number of daily Active cases, new cases, recovered, and deaths of COVID-19 in Indonesia from May 11, 2022, to June 11, 2022 (number of observations 152 days) is shown in Figure 1. Based on the time series data plot of the number of active COVID-19 cases, the time series data is not stationary, with the ADF hypothesis test obtained P-value = 0.4848 which means the data is not stationary concerning the average, see **Table 1**. Furthermore, the logarithm transformation is carried out on the data to stabilize the variance, then the process of differencing on the logarithmic ln (y) data is carried out to stationary the ln(y) data. The results of differencing twice and based on the ADF hypothesis test, p-value = 0.0000 means that log(y) is stationary at level d = 2, see Table 2, thus the estimation of d in the ARIMA model (p, d, q) is obtained d = 2, so the model is ARIMA(p, 2, q).

Figure 2 identification of ARIMA Model (p, 2, q), namely determining p and q with ACF and PACF correlogram plots on the second difference ln(y). Based on the correlogram plot, it is possible that p = 0, 1, 2, 3, and 4 while q = 0, 1, 2, 3, and 4, so there are 24 pairs of p and q values. Of the 24 pairs, the values with the smallest AIC and BIC with eight pairs are stated in **Table 3**. Next, ARIMA (3, 2, 0) is selected with AIC = -2.981139 and BIC = -2.880785 which is the smallest among ARIMA models. Table 4 shows that, ARIMA (3, 2, 0) estimates with AR (1), AR (2), and AR (3) coefficients are significant with p-value = 0.0000, while the intercept is not significant with p-value = 0.9849. Therefore, the equation is

$$(1 - 0.41129\,B - 0.532652\,B^2 - 0.223195\,B^3)\; z_t = \varepsilon_t \quad (9)$$

Where z_t is the second difference of lny.

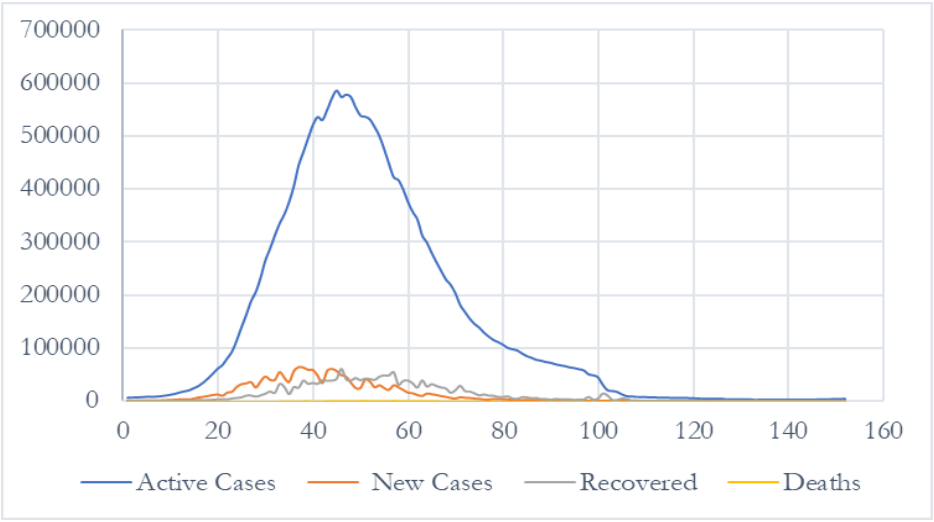

Figure 1. Daily cases of COVID 19 in Indonesia period from January to June 2022.

Table 1. ADF test on logarithm The Active Cases

		t-Statistic	P-value
ADF test statistic		-1.590252	0.4848
Critical values	1%	-3.478189	
	5%	-2.882433	

Table 2. ADF test on The Active Cases logarithm of Second Difference

		t-Statistic	P-value
ADF test statistic		-11.92244	0.0000
Critical values	1%	-3.478189	
	5%	-2.882433	

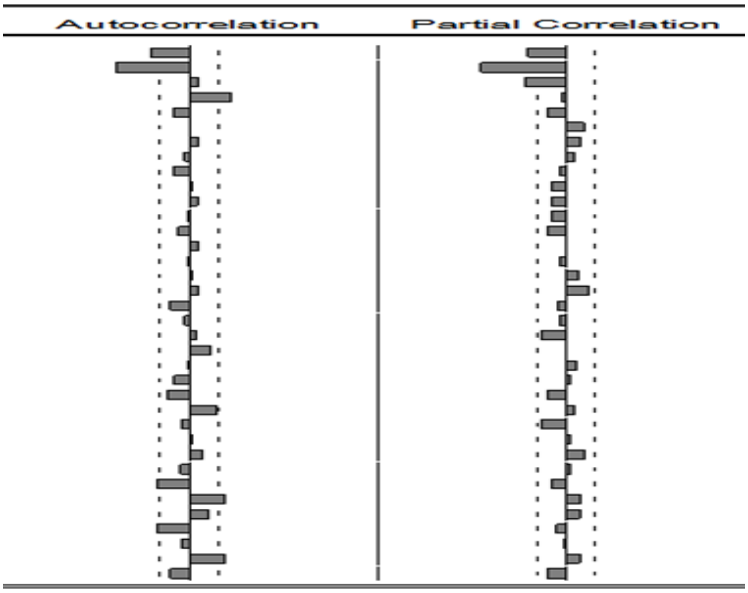

Figure 2. ACF and PACF of the second difference ln(y)

Table 3. Model Selection Criteria

Dependent Variable: DLOG (Y, 2)

Included observations: 150

Model	AIC	BIC
(3,0)	-2.981139	-2.880785
(1,4)	-2.980644	-2.840148
(2,1)	-2.978513	-2.878158
(2,3)	-2.973248	-2.832752
(2,4)	-2.971162	-2.810594
(2,2)	-2.969239	-2.848813
(3,1)	-2.968927	-2.848502
(4,0)	-2.96824	-2.847815

Table 4. Coefficients Estimation of ARIMA(3, 2, 0) Model

Dependent Variable: D (D (LOG(Y)))

Sample: 3 152

Included observations: 150

Convergence achieved after 49 iterations				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.84E-05	0.002548	-0.018993	0.9849
AR(1)	-0.411290	0.055693	-7.385007	0.0000
AR(2)	-0.532652	0.039057	-13.63771	0.0000
AR(3)	-0.223195	0.073889	-3.020698	0.0030

Furthermore, the graph of the number of daily active cases with prediction using equations 9 is shown in **Figure 3**, it can be seen that the predicted value is close to the actual. The next step is MLR modelling with a daily predictor of active cases that depends on daily new cases, recovered and died. By applying equation 8 the estimation of the MLR model is stated in table 5. Based on **Table 5**, the p-value for the estimation of the three regression coefficients is smaller than 0.05 which means that the three estimates are significant, while the *p-value* estimated intercept is greater than 0.05 which means that it is not significant.

The equation model in **Table 5**. is as follows:

$$\text{Active Case} = -869{,}785 + 4{,}026 \times \text{New Case} + 4{,}535 \times \text{Recovered} + 553{,}998 \times \text{Deaths} \quad (10)$$

Furthermore, the results of the analysis of variance, the hypothesis of the model of the number of active cases that depend on the number of new cases, recoveries, and deaths are significantly based on the calculation of a *p-value* less than 0.05 see **Table 6**.

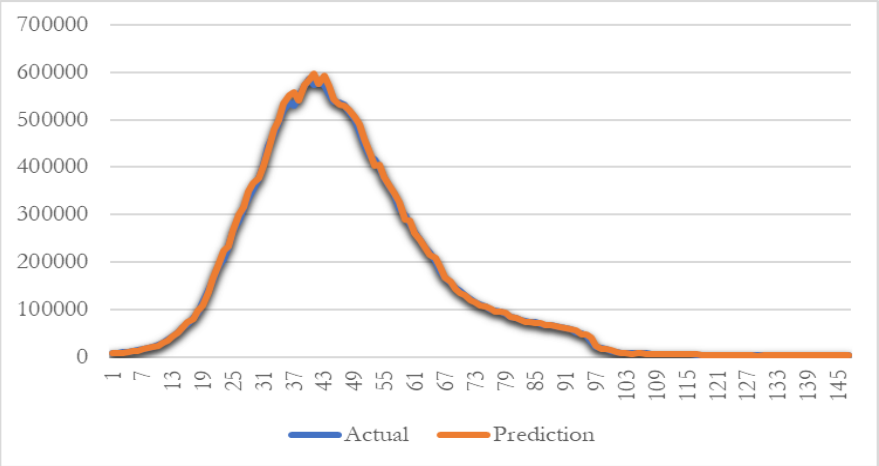

Figure 3. Active case with ARIMA (3,2,0) prediction model

Table 5. Coefficients Estimation of the MLR model

	Coefficients	Standard Error	t Stat	P-value
Intercept	-869.7845431	3757.851512	-0.23145793	0.817279
New Cases	4.02554677	0.250886892	16.04526539	3.35E-34
Recovered	4.53483391	0.691699416	6.556075958	8.64E-10
Deaths	553.9984583	92.98863405	5.957700787	1.79E-08

Table 6. ANOVA of MLR model

	df	SS	MS	F	Significance F
Regression	3	4.96E+12	1.65E+12	1353.713	2.4584E-107
Residual	148	1.807E+11	1.22E+09		
Total	151	5.14E+12			

Prediction of active cases based on the model equation 11 is shown in **Figure 4**.

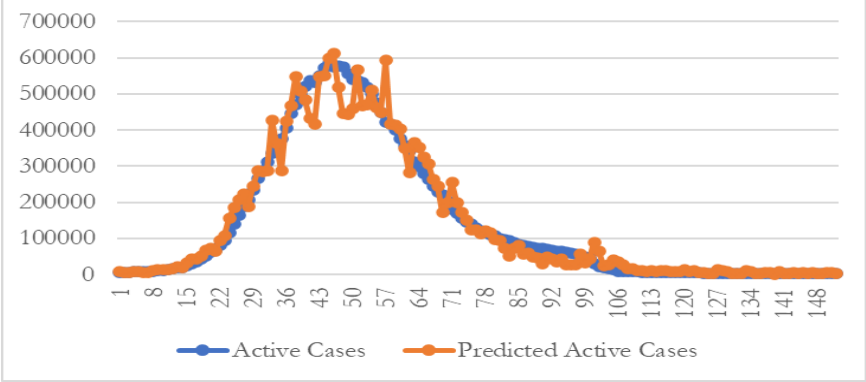

Figure 4. Active case with MLR prediction model

In general, for forecasting purposes, the time series data model in COVID-19 cases uses ARIMA and SARIMA models. This model is still reliable for predicting some pandemic cases, so many researchers use this method to predict pandemic cases. The problem of forecasting is to determine a future value that is close to the actual value. There is no theory that the forecast value will be relative to the actual value so the forecast results may occur far from the actual value. Therefore, many researchers use more than one model for forecasting purposes to obtain the closest forecast value. In this study, the MLR model was used to prove that a month later the prediction of COVID-19 in Indonesia with the ARIMA model was as per the direction of the MLR prediction.

The results of the calculation of the prediction error of the two models the last month from May 12, 2022, to June 11, 2022, are stated in **Table 7**. Based on the error calculation results and the prediction of the number of daily active cases of COVID-19 in Indonesia, the ARIMA model is better than the MLR model. Predictions of the number of active cases from June 12, 2022, to July 11, 2022, both models are shown in **Figure 5**. ARIMA prediction approaches a straight line while

MLR fluctuates around the actual value. Thus, in this case the ARIMA model can be relied upon for daily forecasting a month later. ARIMA and MLR analyzes showed a small increase in active cases. Although this increase is still around 20000 cases throughout Indonesia, we remain vigilant to continue to monitor the increase in active cases.

Table 7. ARIMA and MLR Model Prediction Error

	RMSE	MAE	MAPE
ARIMA	178.663	143.077	3.8636
MLR	3767.42	2606.62	75.1412

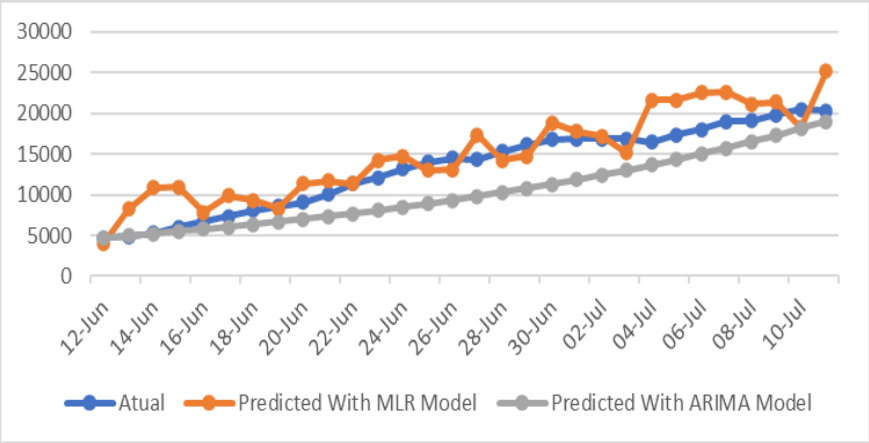

Figure 5. Active Cases for the Period 12 June to 11 July 2022

4. CONCLUSION

The number of active COVID-19 cases in Indonesia has increased again, although not as high as February 2022. The spread of COVID-19 is very dangerous and has many negative impacts, so it requires strict special plans and policies. Controlling and monitoring the development of this pandemic is very necessary, therefore forecasting modeling is a way to monitor the increase in this pandemic. The Time series model for predicting COVID-19 cases is widely used with the ARIMA model, and forecasting the number of future cases is controlled through RMSE, MAE, and MAPE, so some forecasts may deviate far from the actual form. We propose a methodology to consider multiple regression models as a companion to the feasibility of the ARIMA model. The results obtained from forecasting the number of active cases in the direction of the prediction model are not far from the direction of the multiple regression model. The deviation of the forecast with the actual condition is quite small, so the approach used with this method is quite accurate.

AUTHOR'S CONTRIBUTIONS

The authors discussed the results and contributed to from the start to final manuscript.

CONFLICT OF INTEREST

There are no conflicts of interest declared by the authors.

REFERENCES

Abdullah, S., Ismail, M., & Fong, S. Y. (2017). Multiple linear regression (MLR) models for long-term PM10 concentration forecasting during different monsoon seasons. *Journal of Sustainability Science and Management*, *12*(1), 60–69.

AL-Hamad, M. Y., & Qamber, I. S. (2019). GCC electrical long-term peak load forecasting modeling using ANFIS and MLR methods. *Arab Journal of Basic and Applied Sciences*, *26*(1), 269–282.

Ali, M., Prasad, R., Xiang, Y., & Deo, R. C. (2020). Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms. *Renewable and Sustainable Energy Reviews*, *132*, 110003.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*, 105340. https://doi.org/10.1016/j.dib.2020.105340

Bull, S. B. (1998). Regression models for multiple outcomes in large epidemiologic studies. *Statistics in Medicine*, *17*(19), 2179–2197. https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2179::AID-SIM921>3.0.CO;2-L

Bustaman, U. (2021). Mobility-Covid-19 Impact Quadrant : Quantitative Approach To Analyze Community Responses To Covid-19. *Jurnal Aplikasi Statistika & Komputasi Statistik*, *14*(1), 55–68.

Cannon, A. J., & Mckendry, I. G. (1999). Forecasting all-India summer monsoon rainfall using regional circulation principal components: A comparison between neural network and multiple regression models. *International Journal of Climatology*,*19*(14),1561–1578.https://doi.org/10.1002/(SICI)1097-0088(19991130)19:14<1561::AID-JOC434>3.0.CO;2-3

Ceylan, Z, & Bulkan, S. (2018). Forecasting PM10 levels using ANN and MLR: A case study for Sakarya City. *Global Nest Journal*, *20*(2), 281–290.

Ceylan, Zeynep. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of the Total Environment*, *729*. https://doi.org/10.1016/j.scitotenv.2020.138817

Chaurasia, V., & Pal, S. (2020). COVID-19 Pandemic: ARIMA and Regression Model-Based Worldwide Death Cases Predictions. *SN Computer Science*, *1*(5), 1–12. https://doi.org/10.1007/s42979-020-00298-6

Davis, J. K., Gebrehiwot, T., Worku, M., Awoke, W., Mihretie, A., Nekorchuk, D., & Wimberly, M. C. (2019). A genetic algorithm for identifying spatially-varying environmental drivers in a malaria time series model. *Environmental Modelling and Software*, *119*(February), 275–284. https://doi.org/10.1016/j.envsoft.2019.06.010

Ganiny, S., & Nisar, O. (2021). Mathematical modeling and a month ahead forecast of the coronavirus disease 2019 (COVID-19) pandemic: an Indian scenario. *Modeling Earth Systems and Environment*, *7*(1), 29–40. https://doi.org/10.1007/s40808-020-01080-6

Gherghel, I., & Bulai, M. (2020). Is Romania ready to face the novel coronavirus (COVID-19) outbreak? The role of incoming travelers and that of the Romanian diaspora. *Travel Medicine and Infectious Disease*, *34*(February), 101628. https://doi.org/10.1016/j.tmaid.2020.101628

Inyurt, S., Hasanpour Kashani, M., & Sekertekin, A. (2020). Ionospheric TEC forecasting using Gaussian process regression (GPR) and multiple linear regression (MLR) in Turkey. *Astrophysics and Space Science*, *365*(6), 1–17.

Kamboj, V. K., Verma, C., & Gupta, A. (2020). Early Detection of Covid-19 in Canadian Provinces and its Anticipatory Measures for a Medical Emergency. *SN Computer Science*, *1*(6), 1–16. https://doi.org/10.1007/s42979-020-00347-0

Li, J., Deng, D., Zhao, J., Cai, D., Hu, W., Zhang, M., & Huang, Q. (2020). A novel hybrid short-term load forecasting method of the smart grid using MLR and LSTM neural network. *IEEE Transactions on Industrial Informatics*, *17*(4), 2443–2452.

Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(5), 1467–1474. https://doi.org/10.1016/j.dsx.2020.07.045

Suryani, A., & Binarto, A. (2021). Forecasting the Number of New Cases of COVID-19 in Indonesia Using the ARIMA and SARIMA Prediction Models. *2nd International Seminar of Science and Applied Technology (ISSAT 2021)*, 63–68.

WHO.c. (2020). Modes of transmission of the virus causing COVID-19: implications for IPC precaution recommendations. *Geneva: World Health Organization; Available*(March), 1–10. https://doi.org/10.1056/NEJMc2004973.Cheng

Yonar, H. (2020). 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*, *4*(2), 160–165. https://doi.org/10.14744/ejmo.2020.28273

Zhang, S., Diao, M. Y., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *International Journal of Infectious Diseases*, *93*, 201–204. https://doi.org/10.1016/j.ijid.2020.02.033