Advances in Engineering Research, volume 207
Proceedings of the 2nd International Seminar of Science and Applied Technology (ISSAT 2021)

Forecasting the Number of New Cases of COVID-19 in Indonesia Using the ARIMA and SARIMA Prediction Models

Hedi[1,*] Anny Suryani[2] Agus Binarto[3]

[1]*Energy Conversion Engineering Department, Politeknik Negeri Bandung, Indonesia*
[2]*Accounting Department, Politeknik Negeri Bandung, Indonesia*
[3]*Electrical Electronic Engineering Department, Politeknik Negeri Bandung, Indonesia*
**Corresponding author. Email: hedi@polban.ac.id*

ABSTRACT
In June 2020, the Indonesian Government announced to implement a new normal policy as a result of the increasing number of new cases of coronavirus disease (COVID-19) every day, but many new cases until August 2021 were still above June 2020. To control the spread of this pandemic, the Government implements a limiting community activities policy. For this reason, to predict the success of this policy forecasting many new cases in the future is necessary. The purpose of this study is to provide the estimated number of COVID-19 new cases in Indonesia. This study applies two mathematical models: Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA). This research method begins with determining the source of the data. Based on daily observation data from July 25, 2020 to September 9, 2021, identification and estimation of ARIMA and SARIMA modeling were carried out. Based on the calculation results of Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), ARIMA (5, 1, 4) and SARIMA (2, 1, 2)(0, 1, 1)^7 are the most suitable model. Furthermore, based on the calculation results of the smallest Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percent Error (MAPE), the ARIMA(5, 1, 4) model is the most suitable forecasting model for the number of new COVID-19 cases in Indonesia

***Keywords**: COVID-19, ARIMA, SARIMA, Forecasting.*

1. INTRODUCTION

The number of COVID-19 new cases in Indonesia was trending up every day from August 2020 to January 2021 (see Figure 2). In mid-January 2021, there was a very high spike, but then it declined until the beginning of May 2021 with new fluctuating from 3,000 to 7,000 new cases. However, The number of new cases suddenly went up higher than before, starting from the end of May 2021 to July 2021and the peak occurred on July 15, 2021, reaching 56,000 cases. After implementing the policy community activities restrictions, the daily new cases decreased until September 2021 and fluctuating around 22,000 to 30,000. This decline is still very high compared to the decline in the first period. The number of new cases today will simultaneously affect the number of new cases tomorrow, the day after tomorrow and so on. The possibility of reducing spreading cases is social distancing, a clean environment, and not gathering in public places [1]. Therefore a model that can predict the spread of this pandemic is badly needed.

The data of new COVID-19 cases recorded daily is time-series data so that they can be modeled using the ARIMA forecasting model [2] and [3]. By applying this model, it is expected that the number of new cases in Indonesia can be well predicted. Several researchers have applied ARIMA modeling to predict the number of confirmed cases of the COVID-19 pandemic [4], [5], and [6]. This model has a better result compared to the exponential forecasting model [6].

Another model that considers factors in the time series data is SARIMA. Several researchers claimed that the SARIMA model is better than the ARIMA model [7], [8], [9], and [10]. The model of COVID-19 cases generally analyzes data on the number of confirmed cases. By comparing the ARIMA and SARIMA models, this study aims to find the best method for predicting the

daily number of COVID-19 new cases from 2020 to 2021 in Indonesia.

1.1. Related Work

Covid-19 studies have mostly modeled forecasting the number of confirmed cases. One of them is the study conducted by Yonar et al.; they applied the ARIMA model to predict the number of confirmed cases of COVID-19 in several countries, namely, Germany, France, and Turkey [5]. Arun Kumar et al. used ARIMA and SARIMA models to predict the confirmed-cumulative number of COVID-19 cases in the top 16 countries with the most number of global cumulative cases [4].

1.2. Our Contribution

This study aims to find the best forecasting model by comparing the ARIMA and SARIMA. The best model found also shows the implementation impact of the community restriction policy. The findings are expected to predict the number of new cases in Indonesia several months later from September 2021. This forecasting model can detect the possibility of decreasing or increasing the daily number of new cases. This forecasting model may help the Government prepare for the possibility that may occur in the future.

2. MATERIALS AND METHODS

This study used daily data starting from July 25, 2020 to September 9, 2021. The data were analyzed consecutively in five steps: examining data patterns, ARIMA modeling, SARIMA modeling, selecting the best model, and forecasting for the following one month (see Figure 1).

In the first step of the ARIMA model, the data were transformed using logarithms and differencing to stationary concerning variance and mean. Stationary data were carried out with the Augmented Dickey-Fuller (ADF) test.

The second step determined parameters p and q by the autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs. To select the values of p and q, it is necessary to test the fit of the model in terms of explaining the relationship between variables. Information to determine how well the model explains the relationship by applying the AIC and BIC criteria. The smaller the value of AIC and BIC, the better the relationship between variables.

Then the ARIMA (p, d, q) model was obtained with the following equation :

$$\phi_p(B)(1-B)^d(Y_t-\mu)=\theta_q(B)\varepsilon_t$$

and

$$\phi_p(B)=1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p$$
$$\theta_q(B)=1+\theta_1B+\theta_2B^2+\cdots+\theta_qB^q$$

Where : t = 1,2,....T, T = number observed
B = backshift operator
d = differenced level
μ = mean
ε_t= residual

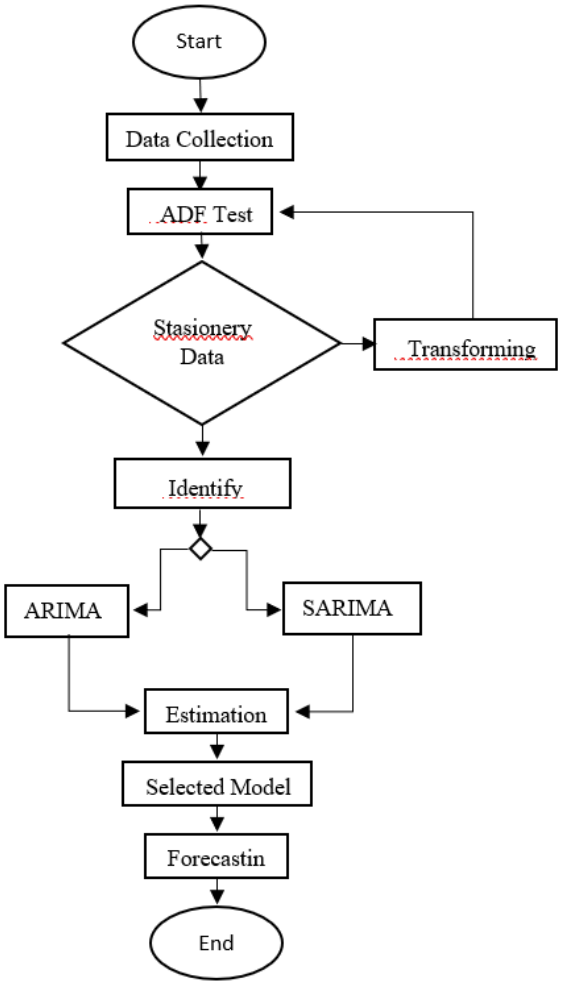

Figure 1 The process of data analysis using ARIMA and SARIMA Models

The SARIMA model is the combination of ARIMA with seasonal elements. The steps were the same as the ARIMA model, namely logarithmic transformation and then transforming to stationary by differencing. ACF and PACF plots were conducted to see the existence of a seasonal as well as to determine the period of S. Then transforming by differencing on S so that the data were stationary. To estimate the P and Q parameters, ACF and PACF plots were performed at the seasonal level. Parameter p, d, P, and Q were determined by applying AIC and BIC to obtain the SARIMA(p, d, q)(P, D, Q)^S model with the equation

$$\phi_p(B)(1-B)^d\Phi_P(B^m)(Y_t-\mu)=\theta_q(B)\Theta_Q(B^m)\varepsilon_t$$

and

$$\Phi_P(B^m)=1-\Phi_1B^m-\Phi_2B^{2m}-\cdots-\Phi_pB^{Pm}$$
$$\Theta_q(B^m)=1+\Theta_1B^m+\Theta_2B^{2m}+\cdots+\Theta_QB^{Qm}$$

The best model was determined based on the calculation results of the RMSE, MAE, and MAPE. The smallest of the three calculations would be the best model.

3. RESULTS AND DISCUSSION

3.1 ARIMA Model

The daily data from July 25, 2020 to September 9, 2021 for 412 days are shown in Figure 2. Graphically the data are not stationary in respect to the variance and mean.

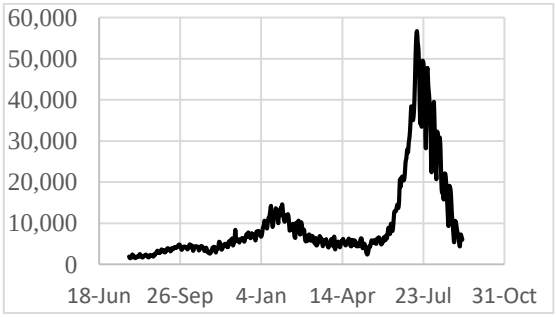

Figure 2 Daily New Cases COVID-19 In Indonesia

These data were transformed by logarithmic transformation to stabilize the variance. Based on the ADF hypothesis test results on the transformed data, the value of t-statistic = -1.961484 is greater than the critical value = -2.868888 with 5% level (see Table 1).This showed that the transformed data ln(y_t) is not stationary in mean.

Table 1. Stationary Test ln(y_t,) using ADF

Null Hypothesis: LNY has a unit root		
		t-Statistic
ADF test statistic		-1.961484
Test critical values:	0.01	-3.447259
	0.05	-2.868888
	0.1	-2.570751

Furthermore, by taking the first differences of the data logarithm and applying the ADF hypothesis test in Table 2, t-statistic = -6.085894 was obtained, and it was smaller than the critical value = -2.8689 with 5% level (see Table 2). This indicated that the data Dln(y_t) were stationary regarding the average, so d = 1 was estimated.

Table 2. Stationary Test first difference ln(y_t,) Using ADF

Null Hypothesis: D(LNY) has a unit root		
		t-Statistic
ADF test statistic		-6.085894
Test critical values:	0.01	-3.447259
	0.05	-2.868888
	0.1	-2.570751

The plots of ACF and PACF patterns on the first differencing data determine the selected models where p, and q parameters (5, 1, 5), (5, 1, 4), (5, 1, 3), (4, 1, 5), (4, 1, 2) , (4, 1, 4), (5, 1, 2), (4, 1, 3), (2, 1, 5), and (2, 1, 3) is shown in Figure 3.

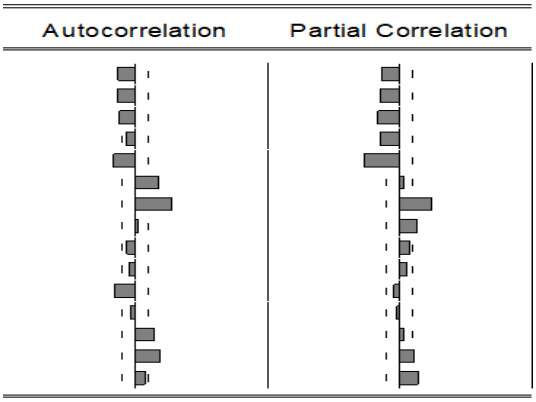

Figure 3 Correlogram of First Difference ln(y_t)

The results of the AIC and BIC calculations on the four selected models are presented in Table 3.

Table 3. AIC and BIC Values in the ARIMA Model

ARIMA	AIC	BIC
(5,1,4)	-1.046620	-0.939067
(5,1,5)	-1.041835	-0.924503
(4,1,5)	-1.037943	-0.930389
(4,1,2)	-1.023792	-0.945572
(5,1,3)	-1.023191	-0.925415
(4,1,4)	-1.021420	-0.923644
(5,1,2)	-1.019953	-0.931955
(4,1,3)	-1.019387	-0.931389
(2,1,5)	-1.019269	-0.931271
(2,1,3)	-1.018173	-0.949730

Table 3 shows two models with the smallest AIC and BIC, namely ARIMA(5, 1, 4) and ARIMA(5, 1, 5). The BIC value of the two models is almost the same, while the AIC of the two models is different. Therefore the ARIMA(5, 1, 4) model is selected. Table 4 depicts the estimation of these model variable coefficient.

Table 4. Variable Coefficient Estimation

Dependent Variable: DLOG(Y)		
Variable	Coefficient	t-Statistic
C	0.002768	0.607300
AR(1)	0.129149	1.083544
AR(2)	-0.687126	-5.449424
AR(3)	-0.138828	-1.046792
AR(4)	-0.621223	-8.069814
AR(5)	-0.350119	-6.830264
MA(1)	-0.462139	-3.797352
MA(2)	0.633950	4.145071
MA(3)	-0.067938	-0.462764
MA(4)	0.561349	6.071020
SIGMASQ	0.019319	16.98011

The number of new daily cases with actual conditions started with a maximum peak on July 15, 2021 and declined until September 9, 2021. Then the forecast for the next month from September 10 to October 17 is shown in Figure 4

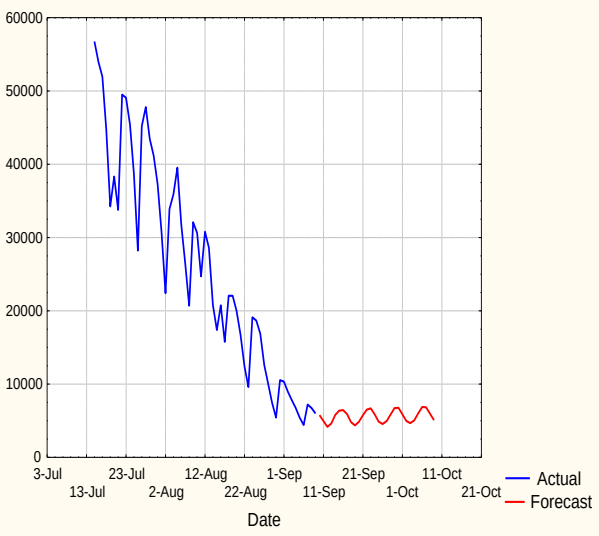

Figure 4 Forecast of ARIMA(5, 1, 4) Model

3.2 SARIMA Model

The identification of the SARIMA model was the same as ARIMA, which was applied to Dln(y_t). Based on Figure 3, there is a seasonal period 7. Then, the first difference is the seasonal part, and based on the ADF test (see Table 5), the t-statistics is smaller than 5% level. This indicated that the first difference of the logarithmic data of the seasonal period 7 is stationary concerning the average.

Table 5. Stationary Test Dln(y_t, 1, 7) Using ADF

Null Hypothesis: D(SY) has a unit root		
		t-Statistic
ADF test statistic		
Test critical values:	1% level	-3.446949
	5% level	-2.868751
	10% level	-2.570678

The next step was the identification. This was to determine the parameters p, q, P and Q, by plotting the ACF and PACF correlograms that die down (see Figure 5). This means that p and q are estimated to be 1, 2, and 3. Furthermore, P = 1, and Q =1, are the values of AIC and BIC with the lowest possible 10 (see Table 6).

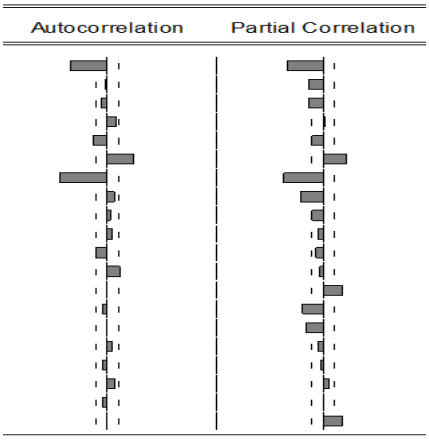

Figure 5 Correlogram of First Difference ln($y_{t,1,7}$)

Table 6. AIC and BIC Values in the SARIMA Model

Model Selection Criteria Table		
Dependent Variable: Dlog(Y, 1, 7)		
Sample: 442		
Included observations: 404		
Model	AIC	BIC
(2,1,2)(0,1,1)	-1.029067	-0.959736
(2,1,2)(1,1,1)	-1.024253	-0.945017
(0,1,2)(0,1,1)	-1.014357	-0.964835
(1,1,1)(0,1,1)	-1.013655	-0.964132
(0,1,1)(0,1,1)	-1.013032	-0.973414
(2,1,1)(0,1,1)	-1.009929	-0.950502
(1,1,2)(0,1,1)	-1.009511	-0.950084
(0,1,2)(1,1,1)	-1.009468	-0.950041
(1,1,1)(1,1,1)	-1.008790	-0.949363
(0,1,1)(1,1,1)	-1.008179	-0.958657

The results of the AIC and BIC calculations of the ten smallest SARIMA models are SARIMA(2, 1, 2)(0, 1, 1)^7. The estimation of variable coeffisien depicted in Table 7.

Table 7. Variable Coefficient Estimation

Variable	Coefficient	t-Statistic
C	-0.000850	-0.903317
AR(1)	-0.323550	-5.807415
AR(2)	-0.171653	-2.802541
MA(1)	-0.040079	-0.786951
MA(2)	-0.002562	-0.060642
MA(7)	-0.776133	-23.04107
SIGMASQ	0.020547	16.92768

Figure 6 illustrates the plot's actual conditions from July 15, 2021 to September 9, 2021. Then the forecast for the next month is from September 10 to October 9, 2021.

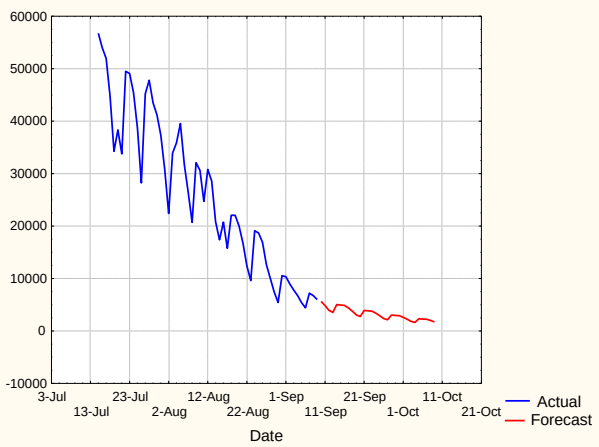

Figure 6 Forecast SARIMA(2, 1, 2)(0, 1, 1)^7 Model

In Table 8 shows the calculation of RMSE, MAE, and MAPE of models ARIMA (5, 1, 4) and SARIMA(2, 1, 2)(0, 1, 1)^7 . The best one is the ARIMA (5, 1, 4) model with the smallest results RMSE = 2178.178, MAE = 1118.790, and MAPE = 10.54833.

Table 8. RMSE, MAE, and MAPE of ARIMA and SARIMA Models

	ARIMA(5, 1, 4)	SARIMA(2, 1, 2)(0, 1, 1)^7
RMSE	2178.178	2296.489
MAE	1118.790	1159.687
MAPE	10.54833	10.7918

The forecast result from both models with observations from September 10, 2021 until October 9, 2021 is shown in Figure 7.

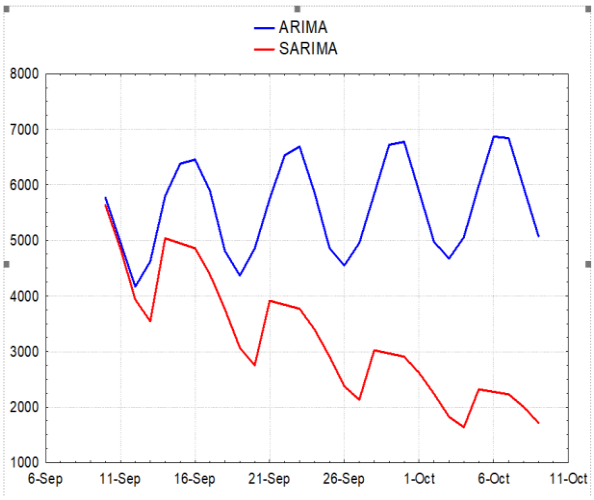

Figure 7 Forecast New Cases Using ARIMA and SARIMA

4. CONCLUSION

The study found that the model for forecasting the number of new cases in Indonesia for a period from July 25, 2020 to September 9, 2021, was the ARIMA (5, 1, 4) and SARIMA (2, 1, 2)(0, 1, 1)^7 models. Graphically the forecasting of the two models is not much different. Considering the results of the RMSE, MAE, and MAPE calculations, the most suitable to predict the number of new cases in Indonesia is ARIMA (5, 1, 4).

The policy of imposing restrictions on community activities implemented by the Government has succeeded in reducing the number of new cases in Indonesia. The forecast indicates this for the next month from September 2021 that it tends to decline.

ARIMA and SARIMA modeling only explain the effect of past data on current data, meaning that the forecasting pattern will follow the pattern that occurred in the past. So this forecasting model can only be used for short-term forecasting. For long-term forecasting, the development of the ARIMA model can be applied, namely the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model.

ACKNOWLEDGMENT

The author would like to thank for the full support from the Energy Conversion Engineering Department, Politeknik Negeri Bandung.

REFERENCES

[1] V. Chaurasia and S. Pal, "COVID-19 Pandemic: ARIMA and Regression Model-Based Worldwide Death Cases Predictions," *SN Comput Sci*, vol. 1, no. 5, p. 288, 2020.

[2] D. Benvenuto, *et. al.*, "Application of the ARIMA model on the COVID-19 epidemic dataset," *Elsevier*, vol. 29, pp. 1 - 4, 2020.

[3] H. Tania Dehesh and P. D. Mardani-Fard, "Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models," *MEdRxiv*, pp. 1-12, 2020.

[4] K. E. ArunKumar, *et. al.*, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA)," *Applied Soft Computing Journal*, vol. 103, pp. 1 - 26, 2021.

[5] H. Yonar, *et. al.*, "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*, vol. 4, no. 2, pp. 160 - 165, 2020.

[6] Hedi and J. Merawati, "Modeling and Forcasting of COVID-19 Confirmed Cases in Indonesia using ARIMA and Exponential Smoothing," in *The International Seminar of Science and Applied Technology*, Bandung, 2020.

[7] A. J. L. Rivero, *et. al.*, "Network Traffic Modeling in a WiFi System with Intelligent Soil Moisture Sensors (WSN) Using IoT Applications for Potato Crops and ARIMA and SARIMA Time Siries," *Applied Science*, October 30 2020 2020.

[8] S. Permana, e*t.al.*, "SARIMA Implementation on Time Series to Forecast the Number of Malaria Incidence," in *International Conference on Information Technology and Electrical Engineering*, Yogyakarta, 2013.

[9] Z. Xinxiang, *et. al.*, "A comparison study of outpatient visits forcasting effect between ARIMA with seasonal index and SARIMA," in *International Conference on Progress in Informatics and Computing (PIC)*, Nanjing, 2017.

[10] M. Valipour, "Long-term runoff study using SARIMA and ARIMA models in the United States," *Royal Meteorological Society*, vol. 22, no. 3, pp. 592 - 598, 2015.