Vol. 3 No. (1) (June 2025) 1-10

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

Prediction of Covid-19 Cases in Indonesia Using the ARIMA Method

Asriyanti Sawiah Adam[1], **Rahma Safira**[2], **M Al Haris**[3], **Saeful Amri**[4]
[1234]Universitas Muhammadiyah Semarang

DOI: https://doi.org/10.26714/jodi.v3i1.212

Article Information

Article History:
Submitted: 20th June 2024
Revised: 23th June 2025
Accepted: 30th June 2025

Keywords:
COVID-19; ARIMA; prediction; spread; Indonesia.

Abstract

This study discusses the use of the ARIMA (Auto Regressive Integrated Moving Average) model to predict the number of COVID-19 cases in Indonesia based on previous data. The results of the analysis show that the ARIMA (1,0,0) model is the most accurate in predicting the spread of COVID-19. Based on this model, the prediction results obtained that confirmed COVID-19 data from January to December 2022 are predicted to decrease. The number of confirmed cases of COVID-19 until December 2022 is predicted to reach 20,0365 cases of spread. So this Covid-19 case still needs special and more serious attention from the government and the public must still be vigilant because based on the results of the study there have been no signs of a significant decrease in the spread of Covid-19 cases. This study provides important insights for the government, medical personnel, and the public in planning strategies for preventing and handling the pandemic.

✉ Corresponding Author:
E-mail: asriyanti588@gmail.com

e-ISSN: 2988 - 2109

INTRODUCTION

Covid-19 virus disease is a new name given by the World Health Organization (WHO) for patients with the 2019 novel corona virus infection which was first reported from the city of Wuhan, China in late 2019. The spread occurred rapidly and created a new pandemic threat. On January 10, 2020, the etiology of this disease was definitely known, namely included in the ribonucleid acid (RNA) virus, namely a new type of corona virus, betacorona virus and one group with the corona virus that causes Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS Cov). The COVID-19 pandemic has presented a major challenge for Indonesia and the world as a whole. Since the SARS-CoV-2 virus was first detected, its spread has had a wide impact on society, health, and the economy. Since the first case was announced in Indonesia in early 2020, the country has struggled to control the spread of the virus and protect public health.

In the face of this complex pandemic, the role of COVID-19 case prediction is becoming increasingly important. Accurate predictions of the number of cases in the future can provide valuable insights for the government, medical personnel, and the public to develop strategies for preventing and handling the pandemic. One method that has proven effective in predicting time series data such as COVID-19 cases is the ARIMA (Autoregressive Integrated Moving Average) method. The ARIMA method is a statistical approach used to forecast time series data by utilizing a combination of autoregressive (AR), integrated (I), and moving average (MA) components. The AR component refers to the dependence of a value on previous values, while the I component reflects the differencing process to make the data stationary. Meanwhile, the MA component relates to the relationship between the current value and the residual values of the AR model. By combining these three components, the ARIMA method can produce accurate predictions of future COVID-19 cases.

Through the prediction of COVID-19 cases using the ARIMA method, it is expected that trends and patterns in previous case data can be identified. This information can help reveal epidemiological seasonality, daily fluctuations, or spikes in cases that may occur in the future. Thus, authorities can formulate more effective and proactive strategies in controlling the spread of the virus and reducing its impact on the community. In addition, the application of the ARIMA method in predicting COVID-19 cases can also provide guidance on resource allocation. By anticipating a possible spike in cases, authorities can better prepare health facilities and medical equipment. This will ensure the readiness of the health system in dealing with a spike in cases and provide adequate care for COVID-19 patients.

Previous studies have shown that the ARIMA method is effective in forecasting COVID-19 cases in several countries. However, the different epidemiological characteristics and public health interventions in each country make it important to develop a forecasting model that is specific to Indonesia. By considering the unique factors in the Indonesian context, the prediction of COVID-19 cases using the ARIMA method is expected to provide more relevant and accurate results for this country.

Factors Causing Covid-19

COVID-19 is caused by the SARS-CoV-2 virus which is transmitted through close contact with an infected person. The main factors causing the spread of COVID-19 include:

1. Human to Human Transmission

COVID-19 is caused by the SARS-CoV-2 virus, which is spread through close contact with an infected person. The virus spreads through respiratory droplets produced when an infected person coughs, sneezes, or talks. When an infected person exhales respiratory droplets, the virus particles can be inhaled by others who are in close proximity. The SARS-CoV-2 virus can also stick to surfaces and objects. People who touch a contaminated surface and then touch their face, especially their eyes, nose, or mouth, can become infected. Human-to-human transmission is a major factor in the rapid and widespread spread of COVID-19 in communities.

2. Asymptomatic or Presymptomatic

One of the unique characteristics of COVID-19 is the ability of the SARS-CoV-2 virus to spread even before a person shows symptoms. Some infected people can be asymptomatic, meaning they do not experience symptoms, but can still spread the virus to others. In addition, there are also people who are presymptomatic, meaning they do not yet show symptoms, but will develop symptoms in the next few days. During this asymptomatic or presymptomatic period, people who are unaware that they have been infected still interact with others and potentially spread the virus without realizing it. This complicates efforts to track and control the spread of COVID-19 because transmission can occur from people who appear healthy and have no symptoms.

ARIMA Model

1. Arima

Auto Regressive Integrated Moving Average or ARIMA is also known as the Box-Jenkins Model (Montgomery & Woodall, 2008). ARIMA is a type of modeling that explains a certain time series based on its own past values (its own lag) and lag estimation errors, so that its equation can be used to predict future values. There are three iteration stages in performing this ARIMA modeling, namely:

1. Determination of a tentative model (model specification) based on sample data to identify the values of p, d, q.
2. The identified parameters of the ARIMA (p,d,q) model are estimated, namely the values of ϕ, Θ, $\sigma 2e$.
3. Analyze the results to see the feasibility of the model. The final model notation is in the form of ARIMA (p,d,q), (Fattah et al., 2018) with

1. *p*is the order for the autoregressive (AR) process
2. *d*is an order that states the number of differentiation processes carried out on time series data that is not yet stationary
3. *q*is the order for the moving average (MA) process

Parameter determination in ARIMA is done by looking at the autocorrelation function or called the Auto Correlation Function (ACF) and the partial autocorrelation function Partial Correlation Function (PACF), as well as differencing. The following is an example of an ARIMA (1,1,1) model consisting of: (Makridakis et al., 1999a).

1. AR Model (1)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$
$$Y_t = \beta_0 - \beta_1 Y_{t-1} = \beta_0 + \epsilon_t$$

From this equation, if a backward shift is performed then:

$$Y_t - \beta_1 B Y_t = \beta_0 + \epsilon_t$$
$$(1 - \beta_1 B) Y_t = \beta_0 + \epsilon_t$$

2. MA Model (1)

$$Y_t = a_0 + e_t - a_1 e_{t-1}$$

and if the equation is shifted backwards (Backward Shift) it becomes:

$$Y_t = a_0 + e_t - a_1 B e_t$$
$$Y_t = a_0 + (1 - a_1 B) e_t.$$

3. *Differentiation*

Differentiating is a process of differentiation that aims to make the average in a time series stationary. The process is written as follows:

$$Y_t' = Y_t - Y_{t-1}$$
$$Y_t' = Y_t - BY_t$$
$$(1 - B) Y_1$$

In the AR(1) and MA(1) models there are 2 constants, namely, whereas in the ARIMA (1,1,1) model there is a constant, assume , then the equation for ARIMA(1,1,1) is:

$$(1 - B)(1 - \beta_1 B) Y_t = \mu + e_t - a_1 B e_t.$$
$$(1 - B)(1 - \beta_1 B) Y_t = \mu - a_1 B e_t + e_t$$
$$(1 - B)(Y_t - \beta_1 B Y_t) = \mu - a_1 B e_t + e_t$$
$$Y_t - \beta_1 B Y_t - B Y_t + \beta_1 B^2 Y_t = \mu - a_1 B e_t + e_t$$
$$Y_t = \beta_1 B Y_t + B Y_t - \beta_1 B^2 Y_t + \mu - a_1 B e_t + e_t$$
$$Y_t = \beta_1 Y_{t-1} + Y_{t-1} - \beta_1 Y_{t-2} + \mu - a_1 e_{t-1} + \epsilon_t.$$

2. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the magnitude of the level of error in the prediction results, where the smaller (closer to 0) the RMSE value, the more accurate the prediction results will be (Sulaiman and Juarna, 2021). RMSE can be used to find out how much error there is in the data from the model used. RMSE can be used as an indicator of incompatibility in modeling. RMSE can be searched

by using:

$$RMSE = \sqrt{\frac{\sum_{i}^{n}(\hat{x} - x_i)^2}{n}}$$

Dimana :

$\hat{x}$ = nilai hasil prediksi

x_i = nilai observasi ke-i

n = banyaknya data

3. **Stationarity**

   The most common approach to creating a stationary series is to reduce previous value from the current value (Khan & Gupta, 2020). Usually denoted with d. If the data is stationary, then d=0.

4. **Validation and Evaluation**

   The model built from training data (70% of data) is used to predict test data (30% of data). The evaluation criteria used are Mean Absolute Error (MAE) (Yang et al., 2020). The MAE formula is as follows (Liang et al., 2020)

$$MAE = \frac{1}{n} \Sigma_{i=1}^{n} |x_i - \bar{x}_i|$$

   *xi*is the actual value of the datais $\bar{x}_i$ the predicted value on the i-th data. n is the number of test data. The smaller the value, the better the model.

METHODS

Data source

The data used in this study is secondary data, namely Marisehat data provided by PT. Maribelajar Indonesia Cerdas as Capstone Project material.

Research Variables

Research variables are basically anything in any form that is determined by the researcher to be studied so that information about it is obtained, then conclusions are drawn. (Sugiyono, 2011) The variables used in this study are:

1. Number of covid-19 cases
2. Time

Research Procedures

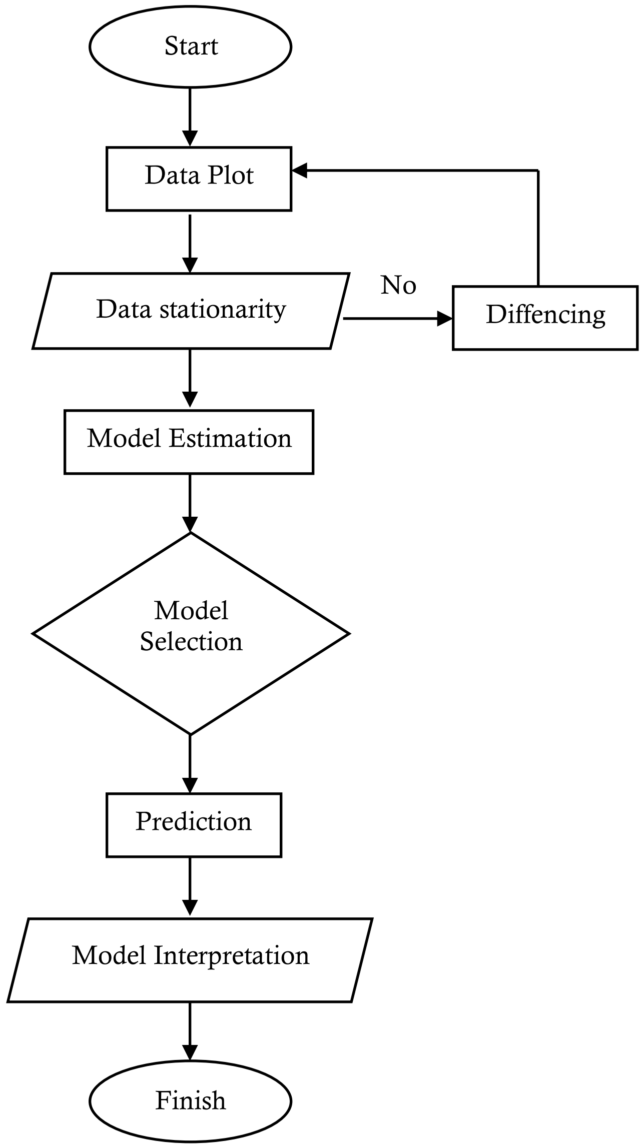


Figure 1. Research Flow

RESULTS AND DISCUSSION

Plot of Active Case Data

The test data has a horizontal pattern (Chart 1). There was a spike in the number of positive COVID-19 cases in Indonesia between May and October in 2020, which resulted in an increase in the need for health facilities and infrastructure.

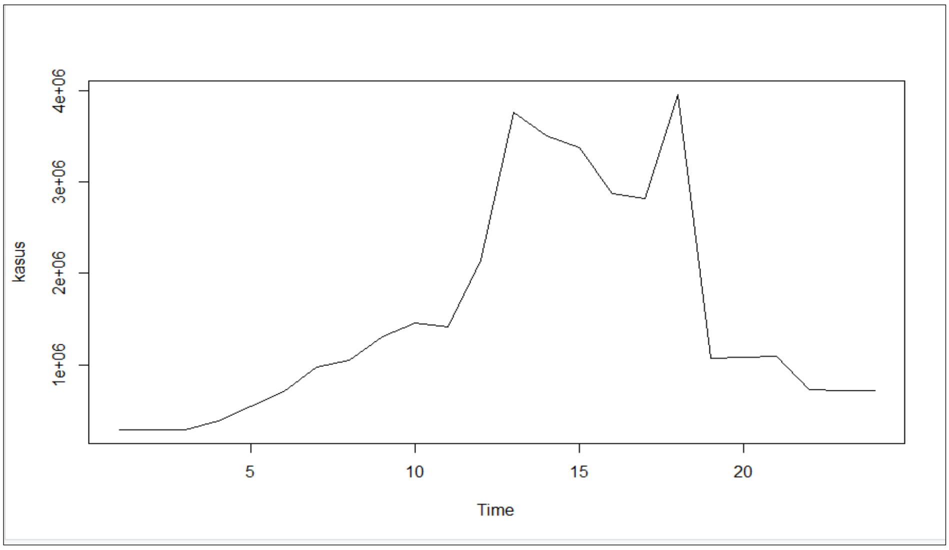
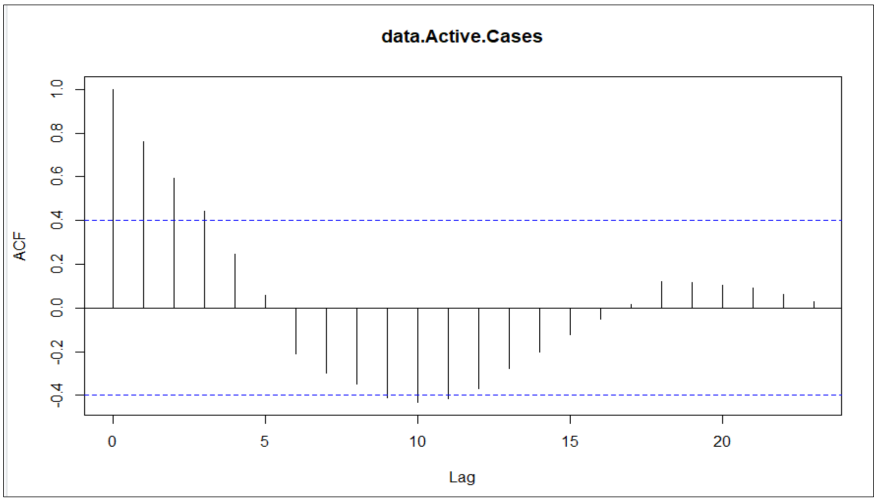


Figure 2. Plot of research data

The initial data (Graph 1) is transformed by Box-Cox (x-2) once to make it stationary in variance. Stationary Time Series data is indicated by the Rounded Value (λ) = 1 (Graph 2). The transformed data is checked for mean stationarity through the autocorrelation function graph (Graph 3) and the partial autocorrelation function (Graph 4). The data is stationary if all its lags are on the Confident Interval (CI) line.

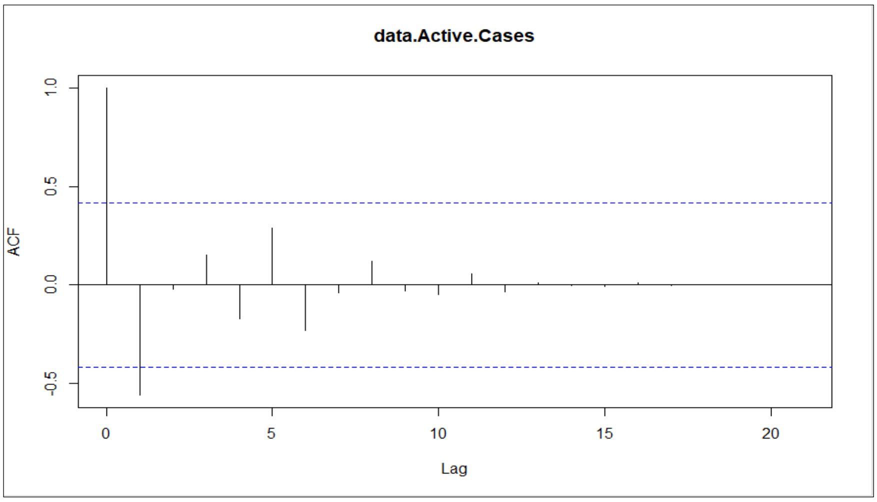
The differencing treatment is carried out because there are 4 lags outside the CI line (Graph 2). Differentencing is only carried out 2 times (d = 2) because the data is already stationary. Furthermore, the autocorrelation function (ACF) graph is used to determine the AR parameters, and the partial autocorrelation function graph is used to determine the MA parameters. Based on the analysis results, several parameters for ARIM are obtained, namely:

1. ARIMA(0,1,1)
2. ARIMA(1,1,0)
3. ARIMA (1,1,1)



ACF with differencing=1

Figure 3. ACF and PACF analysis graph



ACF with differencing=2

Figure 4. ACF function with d=2

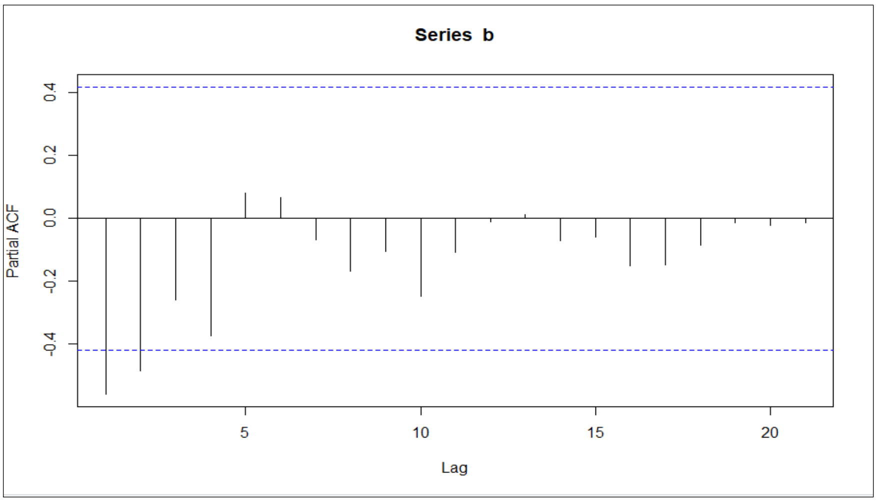


Figure 5. PACF function with d=2

The four ARIMA models produce the following RMSE values:

Table 1. Root Mean Square Error (RMSE) Value of ARIMA Model

Model	Parameter	RMSE
(1,0,1)	AR(1) (MA(1)	4761706
(2,1,2)	AR(2) (MA(2)	4934888
(0,0,1)	MA(1)	6861212
(1,0,0)	AR(1)	2653116

Best Model

From table 1 there are 4 models used to predict the number of positive cases of Covid-19, based on the theory the smaller the MAPE value produced, the better the model. So it is known that the parameters p = 1, d = 0 and q = 0 or ARIMA (1,0,0) with the smallest RMSE value of 2653116 can be used to predict the number of annual positive cases of Covid-19.

Prediction of COVID-19 Spread Cases in Indonesia Using ARIMA (1,0,0)

Table 2. Predicted Data on the Spread of Covid-19 Cases in January 2022 – December 2022.

Month	Confirmed Case Prediction
January 2022	65.0979
February 2022	58.4850
March 2022	52.5439
April 2022	47.2063
May 2022	42.4109
June 2022	38.1027
July 2022	34.2321
August 2022	30.7546
September 2022	27.6305
October 2022	24.8237
November 2022	22.3020
December 2022	20.0365

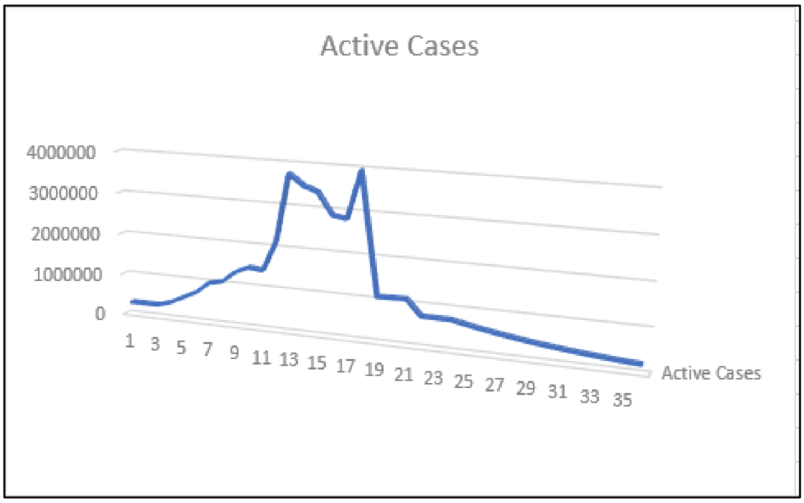


Figure 6. Prediction of Covid-19 cases from January to December in Indonesia

CONCLUSION

Based on data from 2020 to 2021, it is known that positive confirmed data in Indonesia has increased over time. In this study, the best model in predicting cases of the spread of COVID-19 in Indonesia according to the smallest RMSE value is the ARIMA model (1,0,0). Based on this model, the prediction results obtained that confirmed COVID-19 data from January to December 2022 are predicted to decrease. The number of confirmed cases of COVID-19 until December 2022 is predicted to reach 20,0365 cases of spread. So this Covid-19 case still needs special and more serious attention from the government and the public must still be vigilant because based on the results of the study there have been no signs of a significant decrease in the spread of Covid-19 cases.

REFERENCES

[1] Box, G. E,dkk. 2008. "Time Series Analysis: Forecasting and Control. Hoboken": NJ: John Wiley & Sons.

[2] Chatfield, C. 2019. "The Analysis of Time Series: An Introduction (Sixth Edition)": Boca Raton, FL: CRC Press.

[3] Hyndman, R. J., & Athanasopoulos, G. 2018. "Forecasting: Principles and Practice (2nd Edition). OTexts": Melbourne, Australia.

[4] Irawan, A., & Wahyudi, A. 2021. "Time Series Analysis for Forecasting COVID-19 Daily Cases in Indonesia using ARIMA Model".:Journal of Health Informatics and Medical Data Science, 1(1), 45-55.

[5] Kucharski, A.J, dkk. 2020. "Early dynamics of transmission and control of COVID-19": a mathematical modelling study. The Lancet Infectious Diseases, 20(5), 553-558.

[6] Wahyuni Dhiti, dkk. 2021. "Aplikasi Model ARIMA dalam Memprediksi Jumlah Kasus Penyebaran Covid-19 di Provinsi Kepulauan Bangka Belitung":Seminar Nasional Penelitian dan Pengabdian pada Masyarakat.

[7] Rachmawati.A.K. & Misary.S.D. 2021. "Peramalan Penyebaran Jumlah Kasus Virus Covid-19 Provinsi Jawa Tengah dengan Metode ARIMA":Zeta -Math Journal.

[8] Qomariasih Nurul. 2021. "Peramalan Kasus Covid di DKI Jakarta dengan Model ARIMA": Jurnal Syntax Transformasion.