Indonesian Journal of Life Sciences Vol. 02 | Number 02 | September (2020)
http://journal.i3l.ac.id/ojs/index.php/IJLS/

RESEARCH ARTICLE

ARI(1,1) Model for Predicting Covid19 in Indonesia

Nanda Rizqia Pradana Ratnasari

Bioinformatics Department, Indonesia International Institute for Life Sciences, Jakarta, Indonesia

**corresponding author. Email: nanda.ratnasari@i3l.ac.id*

ABSTRACT

Covid19 modelling is needed to help people understanding the distribution or pattern of the data and doing the prediction. The data used for modelling in this study was 'confirmed cases' of Covid19 in Indonesia recorded from March 2 to August 23, 2020. The data was published at Indonesian National Board for Disaster Management - Indonesia Task Force for Covid-19 Rapid Response website. Rstudio software with timeseries package was used for analysing the data. Model obtained from analysis is ARI(1,1) with estimated parameter 0.9859 and standard error 0.0114. Maximum Likelihood was the method conducted to estimate the parameters. The model is appropriate to predict the actual data of Covid19 confirmed cases in Indonesia as the prediction and the actual data plots have a similar pattern.

Keywords: Covid19, autoregressive, prediction.

INTRODUCTION

In early December, Covid19 expanded in Wuhan and exported to growing numbers of countries. The confirmed cases arise with high number of death and significant community transmission occurs in several nations worldwide. However, it becomes a tough problem to quantify the exact numbers of this pandemic as there are not global standards to counter the disease [1], [2]. The disease has brought many new challenges to public health in various nations and it takes decades to recover all these matters [3]. Government of most countries implement new rules to prevent the increasing number of cases or spreading of news that can affect and worsen the situation in society. Individually or in groups, scientists and researchers have conducted some study and investigations in order to flattening the curve for Covid19 [4].

Indonesia, with huge number of populations, is predicted to face difficulties and threat during the pandemic. High mobility and low level of testing become reasons why the positive cases spread vastly and underreported [5], [6]. Moreover, undetected cases caused by unpresented symptoms make the number of patients increase and the number of mortality inaccurate [7]. Indonesian government has conducted some actions to respond the pandemic by applying new regulations such as forming Indonesia Task Force for Covid-19 Rapid Response and by adjusting some rule related to travelling and people mobilization [3]. However, the strategies could not reveals how actually pattern of the confirmed cases in the country [5].

Therefore, to overcome the issues, this paper will provide an approach to measure

the pattern of confirmed cases in Indonesia. This study presents a simple model that can be an alternative for predicting the actual cases in Indonesia. The Autoregressive (AR) Model is proposed as this is one of the simple models for time series data [8]. Autoregressive model can generally be described as a model for values of variables based on the past values of the same variable [9].

The model for Covid19, in general, can help people to understand distribution of some factors such as number of confirmed cases or death. It is also able to predict peak of the outbreak or forecast when the disease will end [10]. However, the model in this study will be used to predict the distribution of the actual data and how the model can forecast values for some period in the future [8].

METHODOLOGY

Data Set and Data Sources

Data used in this study was taken from website of Indonesian National Board for Disaster Management - Indonesia Task Force for Covid-19 Rapid Response on August 25th, 2020. The data recorded some information (variables) related to Covid19 in Indonesia which are number of confirmed cases, death, recovered patients, examined specimens, etc. It has been daily collected since the first case was found in Indonesia which was on March 2nd, 2020. Variable used for the analysis in this study is the daily number of confirmed cases of Covid19 in Indonesia.

The complete data would be split into a training data used to build a model and a test data used to evaluate the predictive validity of the model [5], [8].

Model Formulation

ARIMA model is theoretically best model for predicting time series data. The model consists of three parts which are Autoregressive, Integrated and Moving Average. It is suitable for univariate time series data. The procedure in creating ARIMA model involves some steps, started from fitting an appropriate model, estimating parameters for the chosen orders and verifying the model [5], [8].

As ARIMA model is suitable for time series data, it is widely used in various research regarding epidemiology and public health. In those areas, time series model such as ARIMA is very beneficial since it has been applied to predict actual and future cases. A lot of forecasting in epidemiology has been conducted in several areas such as influenza or dengue cases [5].

ARIMA model assumes that values of variables in particular time is affected by the values of past observations. While AR(*p*) expresses a linear combination of past events with order *p*, the MA(*q*) uses past errors as the explanatory variables [8]. Integrated is used to create a stationary process obtained by differenced a nonstationary process. The process is developed according to the original *Box-Jenkins* methodology [11].

The formulation for ARIMA(*p*, *d*, *q*) is expressed by these equations:

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \omega_t = \sum_{i=1}^{p} \emptyset_i y_{t-i} \quad (1)$$

$$y_t = \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_p \omega_{t-p} + \omega_t = \sum_{j=1}^{q} \theta_j \omega_{j-q} \quad (2)$$

Where y_t is stationary; $\emptyset_1, \emptyset_2, \dots, \emptyset_p$ are constant for AR model with *p* order, and ω_t is white noise series with zero mean and variance σ_ω^2. The AR(*p*) model predict y_t by the previous events $y_1, y_2, \dots, y_{t-1}$. Meanwhile, the MA(*q*) explains y_t by a linear combination of the *q* white noise $\omega_1, \omega_2, \dots, \omega_{t-q}$.

The combination of autoregressive order *p* and moving average order *q* forms a new model called ARMA(*p*, *q*) with formula:

$$y_t = \delta + \sum_{i=1}^{p} \emptyset_i y_{t-i} + \sum_{j=1}^{q} \theta_j \omega_{j-q} + \omega_t \quad (3)$$

The stationary ARMA(*p, q*) process after being differenced d times is notated by ARIMA(*p, d, q*):

$$\Delta^d y_t = \delta + \emptyset_1 \Delta^d y_{t-1} + \cdots + \emptyset_p \Delta^d y_{t-p} + \theta_1 \omega_{t-1} + \cdots + \theta_p \omega_{t-p} + \omega_t \quad (4)$$

where Δ^d denoted *d*-th difference time series.

Plots of autocorrelation function (ACF) and partial autocorrelation function (PACF) are the main graph to identify the parameters of AR and MA model. The AR(*p*) obtains estimation when ACF exhibits the tendency to lie down quickly whereas the PACF denotes the tendency to show spike. In the contrary, the MA(*q*) obtains information when ACF tends to show spike and the PACF exhibits a lie down lines plot [11].

The time series ARIMA model should be stationary and stochastic sequence with zero mean. If there is potency of nonstationary, the data should be transformed using difference of integrated model [5].

(1) Stationarity

Stationary in time series is interpreted as a condition where the data does not have a trend or seasonality. The white noise of the time series data called stationary if it looks much the same in any point of observations [5], [11]. *Box-Jenkins* method determines that time series data is stationary when all the root of the characteristic equations must fall outside the circle [8].

Dickey-Fuller test and Kwiat-kowski-Phillips-Schmidt-Shin (KPSS) test are methods used to check the stationary of time series data. Those tests can be conducted using R software [5], [8].

(2) Identification

Identification is a step of creating time series model recognized by evaluating the order of autocorrelation function (ACF) and partial autocorrelation function (PACF). Subsequently, order of the model in general is determined from 0 to 2 since the model is expected to be simple as possible. The quality of the statistical model and the prediction of the error is given by AIC (Akaike Information Criterion) [5]. The optimal model order is chosen by the number of model parameters, which minimize the AIC [8].

(3) Estimation and Diagnosis

Among all candidate models, the best and appropriate model is diagnosed using residual errors. The errors represent the difference between actual data and the prediction based on model. The errors are expected to be white noise and recognized using *Ljung-Box* method. White noise error means that the residuals are random and do not have a significant correlation [12].

(4) Forecasting

The final process of creating time series model is to do prediction of the actual data and forecasting of future events. The good model may have prediction data that close enough to the actual observation [5], [8].

Statistical Analysis

The analysis conducted in the study is using Rstudio software with timeseries package. The first step established was defining the stationary of the data, which used Dickey-Fuller test. The series and correlation of the data were plotted to see the data sequences and to identify orders of the model. Estimation of the parameters was using maximum likelihood method that was provided in the software. Model fitting tests were carried out to get estimated parameters and significant level, including the standard error, log-likelihood value, AIC and residual errors.

RESULTS AND DISCUSSION

Descriptive Analysis

As explained in the Data Source section, the data use in this study was **confirmed cases of Covid19** in Indonesia recorded from March 2, in which the first case was found in Depok, to August 23, 2020 when the analysis was conducted. The data was in time series form, therefore, it would be analysed using time series model and assumed has ARIMA(*p*, *d*, *q*) model. There were 175 days of recorded data.

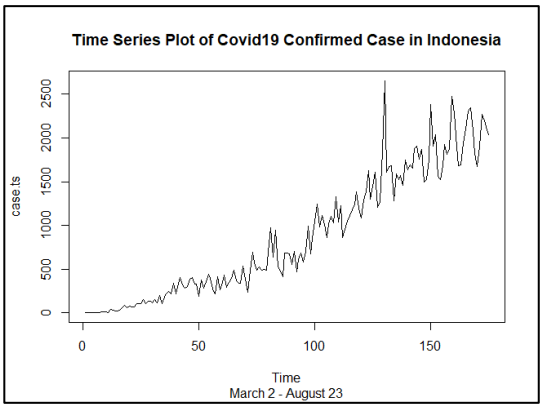

Figure 1. Plot of Covid19 Confirmed Case Data

Figure 1 shows that the time series data of Covid19 confirmed cases has an increasing trend. There is no cyclical or seasonal pattern. The minimum values were zero case that happened on few days at first week of observations. The confirmed case reached the highest number on July 9, 2020 which was 2657. During the pandemic, average value of confirmed case was around 877.

The outlier of the data could not clearly be defined as the data escalated over the time. During day 40-70, the data fluctuated in low number without increasing trend. This might be caused by the factor that during the few first weeks, people tend to obey the policy of self-quarantine. Over the time, especially around day 80-100, there were tendency of inclining trend of confirmed case. This might be caused by policy of interaction restriction.

However, after day 100, the spike of confirmed case number has been high and have tendency to rise over time. Loosening the restriction policy and increasing number of tests became the reasons why the number was getting high.

Model Analysis

The first step in the analysis of time series data is checking the stationary of the data. Dickey-Fuller test was conducted for this purpose.

```
> adf.test(case.ts)

        Augmented Dickey-Fuller Test

data:  case.ts
Dickey-Fuller = -2.9424, Lag order = 5, p-value = 0.1826
alternative hypothesis: stationary
```

Figure 2. Output of Dickey-Fuller Test for Actual Data

Figure 2 represents output from Dickey-Fuller test conducted in Rstudio. The null hypothesis for the test states that the data is nonstationary. The result showed that p-value is higher than 0.05. This indicated that confirmed case is not stationary. Therefore, for further analysis, the data should be transformed using difference.

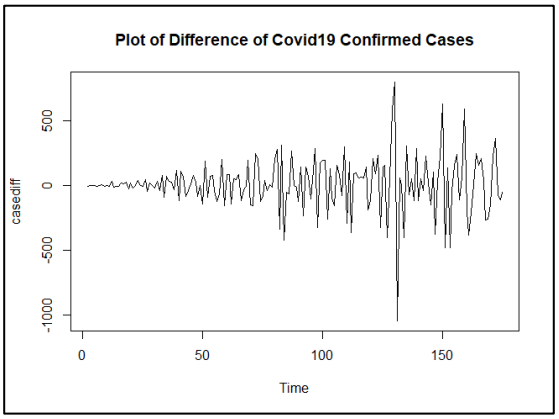

Figure 3. Plot of Difference of Covid19 Confirmed Case Data

```
> adf.test(casediff)

        Augmented Dickey-Fuller Test

data:  casediff
Dickey-Fuller = -11.568, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 4. Output of Dickey-Fuller Test for Differenced Data

The transformed data was plotted and it is showed in Figure 3. The plot showed that the data no longer has a trend. This was confirmed by the Dickey-Fuller test shown in Figure 4 that indicated the data is stationary with p-value of statistics test equals to 0.01 which is less than 0.05. This means that the data could be analysed using integrated *d*=1.

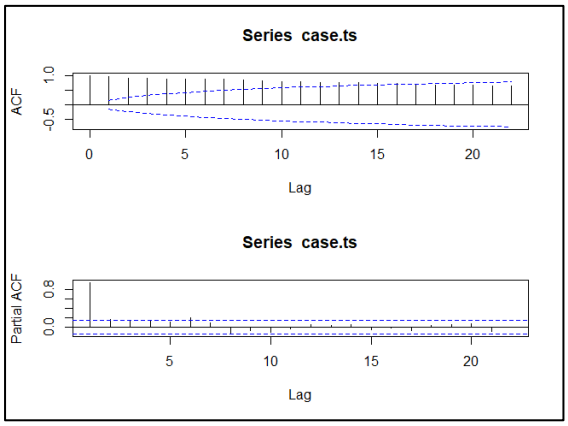

Figure 5. ACF and PACF Plot of Covid19 Confirmed Cases

The order of ARIMA model could be identified using plot ACF and PACF. Figure 5 shows the ACF plots lying down with sinus wave and the PACF plot having tendency to show a spike in the first lag. Based on those plots, the model could be determined using AR model order *p*=1 and integrated order *d*=1, therefore, the model could be recognized as model ARI(1,1).

Based on the model identification, the parameters could be estimated using Maximum Likelihood (ML) method. The parameter of integrated AR(1) for Covid19 confirmed cases was 0.9859 with standard error equal to 0.0114. The AIC obtained from the model estimation was 2367.87.

Figure 6 shows the residual plots calculated from actual data and the fitted model estimation. The residuals had stationary pattern around zero value and it does not significant correlation with other lags. The residuals correlate only with lag=0 which means that it relates to itself.

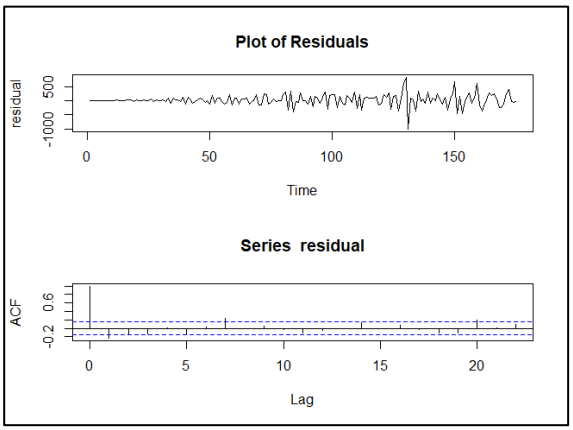

Figure 6. Residuals Plot

Forecasting

Forecasting is needed to check whether the model can represent the actual values. Based on the model identified using ARI(1,1), the prediction of the estimated values can be forecasted. Figure 7 shows that the predicted model could estimate actual data both the pattern and points.

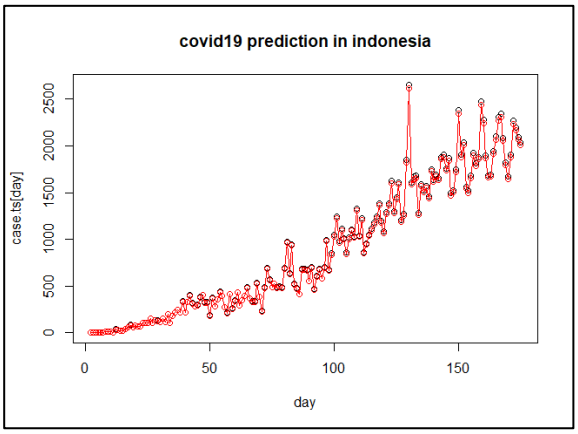

Figure 7. Prediction of Covid19 Confirmed Cases based on model ARI(1,1)

CONCLUSION

Coronavirus disease, known as Covid19, has impact in many aspects of life. The prediction of the confirmed cases can be used to determined policy that may affect society. The data used in this study was confirmed cases of Covid19 in Indonesia recorded from March 2 to August 23, 2020. Rstudio was used to conduct the analysis. Result from the analysis presented that model ARI(1,1) with parameter 0.9859 and standard error 0.0114 was adequate to predict the actual data.

Further analysis is still needed to adjust and fix problems regarding the normality of residuals. Moreover, next steps can be conducted to do forecasting for the future events.

REFERENCES

[1] S. Nadeem, "Coronavirus COVID-19: Available Free Literature Provided by Various Companies, Journals and Organizations around the World," *J. Ongoing Chem. Res.*, vol. 5, no. 1, pp. 7–13, 2020.

[2] F. D. Gennaro, D. Pizzol, C. Marotta, M. Antunes, V. Racalbuto, N. Veronese, and L. Smith, "Coronavirus Diseases (COVID-19) Current Status and Future Perspective: A Narative Review," *Int. J. Environ. Res. Public Health*, vol. 17, no. 8, p. 2690, 2020.

[3] R. Djalante, J. A. Lassa, D. H. E. Setiamarga, C. Mahfud, M. S. Sinapoy, S. Djalante, I. Rafliana, L. A. Gunawano, G. A. K. Surtiari, and H. Warsilah, "Review and analysis of current responses to Covid-19 in Indonesia: Period of January to March 2020," *Prog. Disaster Sci.*, vol. 6, 2020.

[4] N. Nuraini, K. Khairudin, and M. Apri, "Modeling Simulation of COVID-19 in Indonesia based on Early Endemic Data," *Indones. Biomath. Soc.*, vol. 3, no. 1, pp. 1–8, 2020.

[5] F. Fadly and E. Sari, "An Approach to Measure the Death Impact of Covid-19 in Jakarta using Autoregressive Integrated Moving Average (ARIMA).," *Unnes J. Public Heal.*, vol. 9, no. 2, pp. 108–116, 2020.

[6] M. Jefriando and B. C. Munthe, "Indonesia virus death toll rises to highest in Asia outside China," *Reuters*, 2020. .

[7] M. A. Berawi, N. Suwartha, E. Kusrini, A. Herman, Yuwono, R. Harwahyu, E. A. Setiawan, Y. A. Yatmo, P. Atmodiwirjo, Y. T. Zagloel, M. Suryanegara, N. Putra, M. A. Budiyanto, and Y. Whulanza, "Tackling the COVID-19 Pandemic: Managing the Cause, Spread, and Impact," *Int. J. Technol.*, vol. 11, no. 2, pp. 209–214, 2020.

[8] R. Adhikara and R. K. Agrawal, "An Introductory Study on Time Series Modelling and Forecasting," 2013.

[9] T. Saavedra, P. Alvarez de Toledo, A. Crespo Marquez, F. Núñez, C. Usabiaga, and Y. Rebollo, "Autoregressive Models and System Dynamics, A Case Study For The Labor Market In Spain," *Researchgate*, 2020. [Online]. Available: https://www.researchgate.net/publication/228466203_AUTOREGRESSIVE_MODELS_AND_SYSTEM_DYNAMICS_A_CASE_STUDY_FOR_THE_LABOR_MARKET_IN_SPAIN.

[10] A. Agosto and P. Giudici, "A Poisson Autoregressive Model to Understand Covid-19 Contagion Dynamics," *Risk*, vol. 8, no. 77, 2020.

[11] S. Waeto, K. Chuarkham, and A. Intarasit, "Forecasting Time Series Movement Direction with Hybrid Methodolgy. Journal of Probability and Statistics," *J. Probab. Stat.*, vol. 2017, 2017.

[12] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne: OTexts, 2018.