Modeling of covid-19 in Indonesia using vector autoregressive integrated moving average

A Meimela¹,², S S Lestari¹, I F Mahdy¹, T Toharudin¹, and B N Ruchjana³*

¹Department of Statistics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia
²BPS-Statistics of Sumatera Utara Province, Jl. Asrama No 179, Medan 20123, Indonesia
³Department of Mathematics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

*budi.nurani@unpad.ac.id

Abstract. A phenomenon of coronavirus became a big deal around the world at the end of December 2019. To find out how deadly the disease is, we can use the Case Fatality Rate (CFR), which provides the ratio number of deaths due to covid-19 between founded cases number of covid-19. However, studies to see the relationship between the number of cases and the number of deaths caused by covid-19 in Indonesia rarely done. Time Series analysis that can see how the relationship between the number of cases and the number of deaths due to covid-19 in Indonesia is Vector Autoregressive Integrated Moving Average analysis (VARIMA). Data used in this model must be qualified the stationary. For that reason, the transformation using differencing and logarithm on data must be performed to resolve non-stationary. The result shows the model that fulfilled all assumptions and had the smallest AIC value is VARIMA (1,1,1). The model shows the number of cases influenced by the number of cases and the number of deaths in the previous period. The same condition applies to the number of deaths affected by the number of cases from the preceding period.

1. Introduction
At the end of December 2019, the world was shocked by the discovery of the Coronavirus disease, known as covid-19. This disease is caused by a new type of coronavirus (SARS-CoV-2) found in Wuhan, China. Until now, many countries in the world have been infected, including Indonesia. World Health Organization (WHO) finally officially announced that covid-19 had become a pandemic on March 11, 2020 [1]. The first time this disease spread, it often questioned how deadly this disease was and how many people died from this disease [2].

To answer the questions, measuring instrument is used, known as the Case Fatality Rate (CFR). It is the ratio of the number of deaths to the number of founded cases. From this formula, we can assume that when the number of people who died is greater than the number of founded cases [3]. It can conclude that this virus is very deadly, and vice versa. On July 14, 2020, Indonesia's CFR reached 4.7 percent [https://covid19.kemkes.go.id/situasi-infeksi-emerging/info-corona-virus/situasi-terkini-perkembangan-an-coronavirus-disease-covid-19-15-juli-2020/#.XxreI4zBIU]. At the same time, Indonesia's CFR is above the world CFR value of 4.4 percent [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200714-covid-19-sitrep-176.pdf?sfvrsn=d01ce263_2]. This rate figured that this virus is very deadly in Indonesia where the number of death is higher than the
number of case covid-19. That so the government need pay more attention to this case to make the pandemic of covid-19 not being prolonged.

Handling the increase number of the death due to covid-19 is also has to be controlling the increase number of founded cases. To anticipate the increase number of the death, the suspect of covid-19 needs a proper treatment as soon as possible. Research on covid-19 in Indonesia has done a lot, one of the time series analysis on covid-19 is to measure the death impact of covid-19 in Jakarta using ARIMA [4] and the other one is about using the double exponential method on identification covid-19 cases in Indonesia [5]. But the research to see the relationship between the number of covid-19 cases and the number of deaths caused by covid-19 in Indonesia rarely done. Referring to this, it is very important to see the relationship between the number of covid-19 cases and the number of deaths caused by covid-19 in Indonesia and how the relationship between the two.

Time series analysis that can see how the relationship between the number of cases and the number of deaths due to covid-19 in Indonesia is Vector Autoregressive Integrated Moving Average analysis (VARIMA) which explain relationship among observation and error of a variable at certain time with the observation and error of variable previously [6]. VARIMA is easy to apply in multivariate analysis this is one of the benefit from this model [7]. In other research, it is also explained that the VARIMA method can predict more than one variable and explain the relationship between these variables [8]. Therefore, this study will study the relationship between the number of cases and the number of deaths due to covid-19 in Indonesia using VARIMA.

2. Method

2.1. Data Source

The data used in this study are daily data sourced from the worldometers page [https://www.worldometers.info/coronavirus/country/Indonesia/]. The variables used are the number of deaths caused by covid-19 in Indonesia Time series data is a type of data that consists of variables collected according to time sequences within a certain period for a particular category or individual. The selected period time in this study is from 26 May to 14 July 2020. The data processed with R software programming for windows. One of the classifications of time series models is a multivariate time series. It is a time series consisting of several variables that are commonly used to modeling and explain the interactions and movements between some time series variables. A Multivariate time series described in a vector time series [9].

2.2. Stationarity Test

Stationary in time series case means that there are no drastic changes in the data [10]. Data that is non-stationary will give rise to the phenomenon of spurious regression or sharp regression. Spurious regression is a regression figuring the relationship between two or even more variables that appear to be statistically significant when in fact they are not as big as the resulting regression. Augmented Dickey-Fuller (ADF) test is used to see that if data stationary in the mean. Meanwhile, the stationary in the variant is shown through the lambda value from the Box-Cox test.

2.3. Model Identification

The time lag is determined by multivariate looking at the significant lag results on MACF (Matrix Autocorrelation Function) and MPACF (Matrix Partial Autocorrelation Function). MACF or known as the cross-correlation matrix example indicates the structure of the MA (q) model. MACF with n observations can be stated as follows [11]:

\[ \hat{\rho}(k) = [\hat{\rho}_{ij}(k)] \]  (1)

with \(\hat{\rho}_{ij}(k)\) is the sample cross-correlation of the i-th and j-th series components at the time lag k given as follows:

\[ \hat{\rho}_{ij}(k) = \frac{\sum_{t=1}^{n-k}(z_{it}-\bar{z}_i)(z_{jt+k}-\bar{z}_j)}{\left[\sum_{t=1}^{n}(z_{it}-\bar{z}_i)^2\sum_{t=1}^{n}(z_{jt}-\bar{z}_j)^2\right]^{1/2}} \]  (2)
with \( \tilde{Z}_i \) and \( \tilde{Z}_j \) is the sample mean of the corresponding series components.

Meanwhile MPACF or known as the sample partial correlation matrix indicates the structure of the AR model (\( p \)) which is formulated in the following equation [12]:

\[
\phi_{ik} = \frac{\text{cov}\left( (Z_t - \tilde{Z}_t)(Z_{t+k} - \tilde{Z}_{t+k}) \right)}{\sqrt{\text{var}(Z_t - \tilde{Z}_t)} \sqrt{\text{var}(Z_{t+k} - \tilde{Z}_{t+k})}}
\]  

(3)

where \( \tilde{Z}_t \) and \( \tilde{Z}_{t+k} \) are the mean of the minimum squares error in the linear regression estimate of \( Z_t \) and \( Z_{t+k} \) based on the variables \( Z_{t+1}, Z_{t+2}, ..., Z_{t+k-1} \). Partial autocorrelation between \( \tilde{Z}_t \) and \( \tilde{Z}_{t+k} \) equal to the last regression coefficient \( \tilde{Z}_t \) when regressing \( \tilde{Z}_{t+k} \) on the \( k \)-lag variable \( Z_{t+1}, Z_{t+2}, ..., Z_t \) [11].

Criteria for determining the optimum lag length using AICC (Akaike Information Criterion Corrected). Where the optimum lag is at the smallest AICC value. The AICC equation is formulated as follows [12]:

\[
AICC = \log(|\Sigma_{uu}|) + \frac{2r}{(T-r)/k}
\]  

(4)

where \( \Sigma_{uu} = T^{-1} \sum_{t=1}^{T} \tilde{u}_t \tilde{u}_t' \) is the estimated value matrix for the residual covariant of the model using the maximum likelihood estimator, \( k \) is the number of variables, \( r \) is the number of parameters expected and \( T \) is the number of observations.

2.4. Model Vector Autoregressive Integrated Moving Average (VARIMA)

VARIMA is a multivariate time series model which is a development of the Autoregressive Integrated Moving Average (ARIMA) [7]. The advantage of this model is that it can describe data that involves more than one variable. In general, the VARIMA model (\( p, d, q \)) can be written in the following equation [11]:

\[
\Phi_p (B) (I - LB)^d z_t = \Theta_q (B) a_t
\]  

(5)

where:

\( z_t \) : vector observation with \( (z_{1t}, z_{2t}, ..., z_{kt}, ..., z_{kt-1})' \)
\( \Phi_p (B) \) : matrix parameters autoregressive (AR)
\( \Theta_q (B) \) : matrix parameter moving average (MA)
\( (I - LB)^d \): matrix differencing operator
\( a_t \) : random vector with \( a_t \sim MN(0, \Sigma ) \)

2.5. Impulse Response Function (IRF)

Impulse Response Function (IRF) is an analysis that aims at determining how long it takes for a variable to respond to changes that occur in other variables. IRF is also able to trace the contemporary effect of innovation (shock) of a particular variable by one standard deviation on the values of endogenous variables in the system at present and future value [13]. This analysis can be done if all unit-roots of the model formed to have a modulus less than one or are in the unit circle.

3. Result and Discussion

3.1. Research Variable Profiles

The results showed the average number of founded cases and the number of deaths due to covid-19 in Indonesia were 1,117 cases and 47 people, respectively. The standard deviation of each variable is 412.90 and 14.44 (Table 1). Then the highest number of covid-19 cases in Indonesia occurred on July 9, 2020, namely 2,657 cases. The highest number of covid-19 deaths in Indonesia occurred on July 5,
2020, totaling 82 people. Meanwhile, the correlation between the number of cases and the number of death was 58.55%. This value is sufficient so that further analysis can do.

Table 1. Descriptive statistics the number of covid-19 cases and deaths in Indonesia

| Variable           | Mean   | Standard deviation | Maximum | Minimum |
|--------------------|--------|--------------------|---------|---------|
| Number of cases    | 1116.40| 412.90             | 2657    | 415     |
| Number of deaths   | 46.38  | 14.44              | 82      | 22      |

Figure 1. The number of covid-19 cases and deaths in Indonesia

3.2 Stationarity Test
The stationarity test required in this research as the variables used is time-based [https://dx.doi.org/10.2139/ssrn.1911068]. The stationary test can do through the Augmented Dickey-Fuller (ADF) test. ADF test results obtained p-value for the number of cases and deaths of covid-19 of 0.02 and 0.04, respectively. Thus, it concluded that the number of covid-19 cases and death in Indonesia is stationary with an alpha (α) of 0.05. Meanwhile the Box-Cox test, the lambda number of cases is -0.03. So that the Box-Cox transformation required that the variance of data becomes stationary on the other hand, after we do the transformation process, the data are not stationary in the mean. To handle this problem, we need to do a differencing process once.

3.3 Identification Model Vector Autoregressive Integrated Moving Average (VARIMA)
Model identification is multivariate by looking at the smallest AICC value. From table 2, the provisional prediction models are VARIMA (5,1,1), VARIMA (0,1,1), VARIMA (4,1,1) and VARIMA (1,1,1). Of the four prediction models formed, the model that meets all the assumptions is VARIMA (1,1,1).

Table 2. The value of AICC

| Lag | MA 0      | MA 1      | MA 2      | MA 3      | MA 4      | MA 5      |
|-----|-----------|-----------|-----------|-----------|-----------|-----------|
| AR 0| -5.018429 | -5.70749  | -5.596336 | -5.532239 | -5.565358 | -5.566544 |
| AR 1| -5.459321 | -5.676098 | -5.629775 | -5.510133 | -5.388475 | -5.267569 |
| AR 2| -5.496344 | -5.583914 | -5.482659 | -5.428344 | -5.340618 | -5.148184 |
Finally reaches the equilibrium point on the fourth day. Meanwhile, Figure 2 (d) shows that if there is a shock in the number of cases in Indonesia is stationary. Thus, Impulse Response Function (IRF) analysis can do.

### 3.4 Impulse Response Function (IRF)

In Figure 2 the first line shows the response to changes in the number of cases and deaths if there is a shock in the number of covid-19 cases. Where Figure 2(a) shows that the response to changes in the number of covid-19 cases in Indonesia, if there is a shock in the number of covid-19 cases in Indonesia, will decrease over the next one day, then increase until it reaches a balance point on the fourth day. Figure 2(b) shows the response to changes in the number of deaths if a shock occurs in the number of cases, increasing up to one day ahead then decreasing until it reaches a balance point on the second day.

Figure 2 on the second row shows the response to changes in the number of cases and deaths in the event of a shock to the number of covid-19 deaths. Where Figure 2(c) shows that the response to changes in the number of cases if there is a shock, the number of death will fluctuate until the third day until it finally reaches the equilibrium point on the fourth day. Meanwhile, Figure 2(d) shows that if there is a shock in the number of covid-19 cases.
shock in the number of deaths, the response to changes in the number of deaths will reach a balance point on the sixth day.

![Image](image_url)

**Figure 2.** *Impulse response function* the number of covid-19 cases and deaths in Indonesia

4. **Conclusion**

The model formed between the number of founded cases and the number of deaths due to covid-19 in Indonesia is VARIMA (1,1,1). In this model the number of cases and the number of covid-19 deaths in the previous period affect the number of deaths and the number of current covid-19 cases. If there is a shock to the number of cases, it will take about two days for the number of deaths to return to equilibrium. On the other hand, if there is a shock to the number of deaths, it will take up to four days for the number of cases to come to an equilibrium.

**Acknowledgments**

The authors thank Rector Universitas Padjadjaran who provided financial assistance for the dissemination of research through Library and Online Data Research Grant 2020 with contract number: 1735/UN6.3.1/LT/2020. Also thanks to the reviewers for the valuable review of this paper.

**References**

[1] Cucinotta D M V 2020 WHO Declares COVID-19 a Pandemic *Acta Biomed.* 91 157–160
[2] Chang C S, Yeh Y T, Chien T W, Lin J J C, Cheng B W, Kuo S C 2020 The Computation of Case Fatality Rate for Novel Coronavirus (COVID-19) Based on Bayes Theorem: An Observational Study *Medicine.* 99 1-8
[3] Mi Y, Huang T, Zhang J, Qin Q, Gong Y, Liu S, Xue H, Ning C, Cao L, Cao Y 2020 Estimating the Instant Case Fatality Rate of COVID-19 in China Int. J. Infect. Dis. 97 1-6

[4] F. Fadly and E. Sari 2020 An Approach to Measure the Death Impact of Covid-19 in Jakarta using Autoregressive Integrated Moving Average (ARIMA) Unnes J. Public Heal 9 108–116

[5] S. Harini 2020 Identification COVID-19 Cases in Indonesia with The Double Exponential Smoothing Method J. Mat. “MANTIK” 6 66–75

[6] A. Setiawan, M. N. Aidi, I. M. Sumertajaya, 2015 Modelling of Forecasting Monthly Inflation By Using Varima and Gstarima Models Forum Statistik Dan Komputasi 21 60–63

[7] Fitrianti H, Belwawin S M, Riyana M, Amin R 2019 Climate Modeling Using Vector Moving Average Autoregressive IOP Conference Series Earth Environment Science 343 012201

[8] Ayudhiah M P, Bahri S, Fitriyani N 2020 Peramalan Indeks Harga Konsumen Kota Mataram Menggunakan Vector Autoregressive Integrated Moving Average (VARIMA) Eigen Maths J. 3 1-8

[9] Gusnadi R, Rahmawati R, Prahutama A 2015 Pemodelan Generalized Space Time Autoregressive (GSTAR) Pada Data Jumlah Wisatawan Mancanegara Empat Kabupaten/Kota di Jawa Tengah Jurnal Gaussian 4 17-26

[10] Makridakis S, Wheelwright S C, McGee V E 1983 Forecasting Methods and Applications 2nd ed (Canada: John Wiley & Sons,Inc)

[11] Wei W W S 2006 Time Series Analysis Univariate and Multivariate Methods (New York: Addison-Wesley Publishing Company, Inc)

[12] Siregar D, Aidi M N, Sumertajaya I M 2015 Proc. International Conference on Research, Implementation, and Education of Mathematics and Sciences 2015 (Yogyakarta State University) 165-74

[13] Lianti T M 2017 Interkorelasi Antara BI Rate Dengan Bagi Hasil Tabungan Bank Syariah Di Indonesia J. Ekonomi dan Bisnis, 17 58-72

[14] Hannan E J 1970 Multiple Time Series (New York: John Wiley & Sons,Inc)