The autoregressive integrated vector model approach for covid-19 data in Indonesia and Singapore

R Fitriani¹,², W D Revildy¹,³, E Marhamah¹, T Toharudin⁴, and B N Ruchjana*⁵

¹Magister Student of Applied Statistics Universitas Padjadjaran, Jl. Dipati Ukur No.35, Bandung 40132, Indonesia
²Statistics of Blitar Municipality, Jl. Kenari No.62 Blitar 66134, Indonesia
³Statistics of Sorong Selatan Regency, Jl. Teminabuan-Ayamaru Sorong Selatan 98454, Indonesia
⁴Department of Statistics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia
⁵Department of Mathematics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

*budi.nurani@unpad.ac.id

Abstract. Almost all of the countries in the world were attacked by coronavirus disease 2019 (covid-19) which is started from pneumonia endemic cases in Wuhan China in December 2019. The virus spreads including in the Southeast Asian region where Singapore and Indonesia are two countries with the highest cases. Simultaneous modeling of cases in both countries is important as information on the development of covid-19 cases. The Vector Autoregressive Integrated (VARI) model is a multivariate time series model that can be used to build model non-stationary time series data in several locations simultaneously. This study aims to analyze the development of covid-19 cases and builds a VARI model of covid-19 cases in Indonesia and Singapore. The data used from daily covid-19 confirmed cases in the period March 16th until April 19th, 2020. Plotting and statistics testing covid-19 data series from both countries show non-stationary series which have trend and fluctuation. Based on optimum lag identification use the minimum Akaike Information Criteria correction (AICc), the parsimony model is obtained namely VARI(1,1) which has satisfied the multivariate normal, white noise, and homogenous assumption. The result shows that covid-19 cases in Indonesia and Singapore have a strong positive correlation. However, the covid-19 cases in both countries were only influenced by previous cases in each country. The accuracy shows that the model is good enough for forecasting covid-19 cases in both countries.

1. Introduction
At present, the world is rocked by an outbreak starting with a mysterious case of pneumonia in Wuhan China which was reported in December 2019 [1]. The pneumonia case spread and infected people outside of Wuhan, even throughout China. The World Health Organization (WHO) officially announced "Coronavirus Disease (covid-19)" as the name of the new disease on February 12th, 2020 caused by the virus "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)" [2]. The virus spread to other countries then WHO confirmed covid-19 as a pandemic on March 11th, 2020. The covid-19 confirmed cases in the world up to April 19th, 2020, are 2,324,731 cases which spread in 215 countries, with a mortality rate of 160,434 people [14].
The spread of covid-19 also occurred in the Southeast Asia region, with Indonesia and Singapore as the 2 countries with the highest cases. On April 19th, 2020, there were many confirmed cases in Indonesia and Singapore as many as 6,575 and 6,588 cases, respectively. The death rate in Indonesia was 582 people while the death rate in Singapore was 11 people [14].

The impact of the covid-19 pandemic has been felt across countries, communities, and economies. In most areas, the government faces a difficult choice between public safety or reviving the economy [3]. Lockdown and physical distancing policies to prevent the spread of viruses harm the economy of developing countries and are even prone to triggering poverty and inequality [4].

Since the first case of covid-19 was recorded, the number of cases in Indonesia and Singapore until the end of April 2020 shows an increase and it is uncertain when the covid-19 pandemic will come to an end. Research related to covid-19 in Indonesia using time series data, in particular, is still not much done. Previous studies related to covid-19 in Indonesia showed that the trend of covid-19 cases overtime followed an exponential function [5]. Other studies have found that the proximity of a country’s territory affects each other covid-19 case number [6].

An analysis is needed regarding covid-19 cases to obtain a simultaneous model of covid-19 cases in Indonesia and Singapore. Hence, it can give information on the development of covid-19 cases in both countries. The VARI model is a multivariate time series model that can be used to build model non-stationary time series data in several locations simultaneously [7]. Thus, this study aims to analyze the development of covid-19 cases and apply the VARI model of covid-19 cases in Indonesia and Singapore.

2. Method
2.1. Data Source
This study engaged the covid-19 data obtained from the worldometers page (https://www.worldometers.info/coronavirus/). The variable used is the number of daily confirmed cases in Indonesia and Singapore period March 16th until April 19th, 2020. The simultaneous model approach of covid-19 cases in the two countries is the multivariate time series method.

2.2. Stationary
Time series analysis can be performed on stationary conditions. If the observed series have properties that do not depend on the time is called a stationary time series [8]. According [9] a stochastic process is stationary if its first and second moments are time-invariant. A stationary process \( z_t \) have the mean \( E(z_t) = \mu \) and variance \( \text{Var}(z_t) = E(z_t - \mu)^2 = \sigma^2 \), and the covariance \( \text{Cov}(z_t, z_{t+k}) = E(z_t - \mu)(z_{t+k} - \mu) \) [10]. A differencing process is needed when it is non-stationary in the mean and transformations such as logarithms are used to stabilize the variance.

The stationary condition for Vector Autoregressive (VAR) model can be investigated simply. A necessary condition for the VAR(1) series \( z_t \) to be stationary is that all eigenvalues must be less than 1 in absolute value. It can also be shown that if all eigenvalues are less than 1 in absolute value, then the VAR(1) series \( z_t \) is stationary [11].

2.3. Order Identification
Autocorrelation function (ACF) and partial autocorrelation function (PACF) are a tool for identifying the order of univariate AR(\( p \)) and MA(\( q \)) model. For the multivariate time series model, identifying the order is using matrix autocorrelation function (MACF) and matrix partial autocorrelation function (MPACF). The correlation matrix function is defined as \( \tilde{\rho}(k) = \tilde{\rho}_{ij}(k) \), [10] where:

\[
\tilde{\rho}_{ij}(k) = \frac{\sum_{t=1}^{n-k}(z_{t+k-j}^i)(z_{t+k-j}^j)}{\sum_{t=1}^{n}(z_{t}^i - \bar{z}_{i})^2 \sum_{t=1}^{n}(z_{t}^j - \bar{z}_{j})^2}^{1/2},
\]

The partial autoregression matrix at lag \( s \), denoted by \( \varphi(s) \), to be the last matrix coefficient when the data are fitted to a vector autoregressive process of order \( s \) [10]. The partial autoregressive matrix function is defined:

\[
\varphi(s) = \begin{cases} 
\Gamma(1)[\Gamma(0)]^{-1} & s = 1 \\
[\Gamma(s) - c'(s)[A(s)]^{-1}b(s)][\Gamma(0) - b'(s)[A(s)]^{-1}b(s)]^{-1} & s > 1 
\end{cases}
\]
where
\[
A(s) = \begin{bmatrix}
\Gamma(0) & \Gamma(0) & \ldots & \Gamma'(s - 2) \\
\Gamma(0) & \Gamma(0) & \ldots & \Gamma'(s - 3) \\
\vdots & \vdots & \ddots & \vdots \\
\Gamma(s - 2) & \Gamma(s - 3) & \ldots & \Gamma(0)
\end{bmatrix},
\begin{bmatrix}
\Gamma'(s - 1) \\
\Gamma'(s - 2) \\
\vdots \\
\Gamma'(1) \\
\Gamma'(0)
\end{bmatrix},
\begin{bmatrix}
\Gamma(1) \\
\Gamma(2) \\
\vdots \\
\Gamma(s - 1)
\end{bmatrix}.
\]

For the vector model AR(p), the equation (2) can be written:
\[
\phi(s) = \begin{cases}
\phi_p, & s = 1 \\
0, & s > 1
\end{cases}
\]

The matrices of MACF and MPACF that is formed are complex when the dimension of the vector is increased. The simple method for identification of MACF and MPACF is given by [10] using the sign (+), (-), and (.) in the (i,j) position of matrix. Sign (+) denotes a value greater than 2 times estimated standard errors, (-) denotes a value less than -2 times estimated standard errors, and (.) denotes a value within 2 estimated standard errors. The optimum lag determination is fulfilled when the lag (s) is dominated by sign (.). Other alternative to determine the optimum lag uses Akaike Information Criteria (AICc). The optimum lag is determined by the lowest value of AICc. The equation of AICc is [12]:
\[
\text{AICc} = n \left( \log(\hat{\Sigma}) + m \right) + 2b\left(p m^2 + m(m + 1)/2 \right) \text{ with } b = n/[n - (p m + m + 1)]
\]

where \( n \) denotes the number of observations, \( m \) denotes the dimensional autoregressive vector, \( p \) is autoregressive order, and \( \Sigma \) is the variance-covariance matrix.

2.4. Vector Autoregressive Integrated (VARI)
The VARI model is a development of Autoregressive Integrated Model (ARI) which is influenced by the variable itself and other variables in the previous period with non-stationary data [13]. If the data is carried out the process first differencing to produce stationary data, the VAR(1) model becomes VARI(1,1) model with the following equation:
\[
z_t - z_{t-1} = \Phi(z_{t-1} - z_{t-2}) + e_t
\]
where \( z_t \) is the multivariate time series vector of observations at time \( t \), \( \Phi \) is square parameter matrices, error is white noise with mean 0 and variance \( \sigma^2 \), and \( e_t \sim N(0, \sigma^2) \), if \( y_t = z_t - z_{t-1} \) then the equation (6) can be written
\[
y_t((N \times (T-1) \times 1)) = Y^*_t(N \times (T-1) \times (N \times N)) + e_t((N \times (T-1) \times 1))
\]

The estimation parameter of VARI model uses Ordinary Least Squares (OLS) procedures. It is done by simplifying eq.(7) to be a linear model as follows:
\[
y = X\beta + \epsilon
\]

where \( y = y_t((N \times (T-1) \times 1)) = \begin{bmatrix}
y_2((N \times (T-1) \times 1)) \\
y_3((N \times (T-1) \times 1)) \\
\vdots \\
y_{T-1}((N \times (T-1) \times (N \times N)) \end{bmatrix}, X = Y^{t-1}_t((N \times (T-1) \times (N \times N)) = \begin{bmatrix}
y'_2((N \times (T-1) \times (N \times N)) \\
y'_3((N \times (T-1) \times (N \times N)) \\
\vdots \\
y'_{T-1}((N \times (T-1) \times (N \times N)) \end{bmatrix}, \beta = \Phi^*_t(N \times (N \times 1)))
\]

and \( \epsilon = e_t((N \times (T-1) \times 1)) \). The estimation of parameter can be obtained by minimize the sum square error.

2.5. Mean Absolute Percentage Error (MAPE)
The model criteria selected depend on the goodness of fit, such as the residual mean square. However, if the model purpose for forecasting future values then it can be based on forecast errors as the alternative criteria for model selection [10]. The MAPE can be written as follows:
\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{\tilde{z}_t - \hat{z}_t}{\tilde{z}_t} \right| \times 100\%
\]

with \( \tilde{z}_t \) is forecast data.

3. Result and Discussion
3.1. Descriptive Analysis
The first confirmed covid-19 cases in Indonesia were reported on March 2\textsuperscript{nd}, 2020, while Singapore announced the first confirmed cases on January 23\textsuperscript{th}, 2020. The daily confirmed cases in Indonesia and
Singapore continued to increase. Figure 1 shows that in the period March 16th until April 19th, 2020, the confirmed cases in the two countries have the same pattern relatively. The series looks fluctuating and increasing. It could be explained if the confirmed covid-19 cases in one country increase or decrease then it would be accompanied by the increased or decreased confirmed covid-19 cases in the other country.

![Figure 1. The confirmed covid-19 cases in Indonesia and Singapore on March 16th until April 19th, 2020.](image)

Table 1. Description of confirmed covid-19 cases in Indonesia and Singapore on March 16th until April 19th, 2020.

| Country  | Minimum | Mean     | Maximum | Standard deviation |
|----------|---------|----------|---------|--------------------|
| Indonesia| 17      | 184.5143 | 407     | 114.6161           |
| Singapore| 17      | 181.7714 | 942     | 228.749            |

The coefficient correlation confirmed covid-19 cases between Indonesia and Singapore was 0.7842 which shows that there was a quite strong positive relationship confirmed cases between the two countries. The correlation between them means that the built model was one variable comprise of the function of other variables, therefore the appropriate model in this data is a multivariate time series model.

3.2. Stationary

According to Figure 1, it can be seen that the series of the two countries has an element of an increasing trend and tends to fluctuate. This indicates that the two series are not stationary. The stationary testing was carried out using Box-Cox Lambda statistics to test stationary in the variance and Augmented Dicky Fuller (ADF) statistics to test stationary in the mean.

The Box-Cox Lambda statistics were less than one which means that two series not stationary in the variance then the logarithm transformation is carried out. Next, the Indonesia and Singapore series were tested using the ADF test, the results showed that the p-value was 0.5495 and 0.9364, respectively which
was greater than the significance level of $\alpha = 5\%$. This means that the two series were not stationary in the mean and hence first differencing was carried out. After the first differencing on both countries series, the result of the ADF test showed that the p-value was 0.0187 and 0.0206, respectively which was less than the significance level of $\alpha = 5\%$. This means that the two series were stationary in the mean. Figure 2 shows that the two series are already stationary.

![Figure 2](image)

**Figure 2.** Plot the stationary data of confirmed covid-19 cases in Indonesia and Singapore.

### 3.3. Identification Model

The time series model identification is through Partial Autocorrelation Function (PACF) as seen in Figure 3 and Figure 4. Both series Indonesia and Singapore show the PACF plot cut off after lag 1 which indicated that both countries have the same phenomenon model AR(1). The multivariate time series identification can also be done using MACF and MPACF plots. The MACF and MPACF plot are quite accurate but less practical and tend to be subjective, and hence the time series order determination is also done by looking at the AICc values.

![Figure 3](image)

**Figure 3.** Plot PACF Indonesia series.

![Figure 4](image)

**Figure 4.** Plot PACF Singapore series.

Figure 5 and Figure 6 show the MACF and MPACF plots from differencing data. Both the MACF and MPACF plot was signed (-) in the lag 1. It means the negative correlation was statistically significant. Meanwhile, the sign (+) is seen dominate all lags after lag 1. It shows that no correlation occurrence significantly. It can be concluded that there were cut-off after lag 1 in MACF and MPACF consequently, there will be more than one model to be composed. Therefore, identified AICc values in some order is required. The order which has the smallest AICc value is the appropriate VARI model.

![Variable/Lag](image)

**Figure 5.** MACF plot of the confirmed cases covid-19 in Indonesia and Singapore.

![Variable/Lag](image)

**Figure 6.** MPACF plot of the confirmed cases covid-19 in Indonesia and Singapore.

The result of the identification of AICc values for all model's possibility is as follows in Table 2. The smallest AICc values were in the AR(1) model of 17.126 which means the optimum lag was an
autoregressive order 1. Therefore the best multivariate time series model for confirmed covid-19 cases of Indonesia and Singapore series was VARI(1,1). The multivariate stationery can also be detected from the corresponding eigenvalues of the VARI(1,1) model which were $\lambda_1 = 0.4630$ and $\lambda_2 = 0.4194$. Those eigenvalues were satisfied with the stationary theorem which stated if the eigenvalues $|\lambda_i| < 1$ then the multivariate time series model was stationary, and hence the VARI(1,1) model for confirmed covid-19 cases of Indonesia and Singapore was stationary.

**Table 2. AICc value from several model order.**

| Lag | MA 0   | MA 1   | MA 2   | MA 3   | MA 4   | MA 5   |
|-----|--------|--------|--------|--------|--------|--------|
| AR 0| 17.385 | 17.746 | 18.040 | 18.025 | 18.169 | 18.549 |
| AR 1| 17.126 | 17.549 | 17.872 | 17.814 | 18.243 | 18.640 |
| AR 2| 17.202 | 17.513 | 17.914 | 18.119 | 18.435 | 18.954 |
| AR 3| 17.510 | 17.750 | 17.973 | 18.193 | 18.487 | 19.340 |
| AR 4| 17.745 | 17.873 | 18.136 | 18.416 | 18.331 | 18.590 |
| AR 5| 17.784 | 17.781 | 17.743 | 18.101 | 19.100 | 19.505 |

**3.4. VARI(1,1) Model**

Parameter estimation of the VARI(1,1) model was done by the Ordinary Least Squares (OLS) method as seen in Table 3. The results show that the p-value of F test in Indonesia and Singapore series were 0.03022 and 0.03748, respectively which less than the significance level of $\alpha = 5\%$. This means that there was at least one variable that has a significant effect on each model. The partial test results on the two-equation models show the same result that only the variable itself in the previous period has a significant effect. This indicates that the covid-19 cases in Indonesia and Singapore were only affected by the previous covid-19 cases in each country. In other words, there was no influence between countries.

**Table 3. Parameter estimation VARI(1,1) model.**

| Eq. | Variables   | Coefficient | P-value of Partial Test | F Statistics (P-value) |
|-----|-------------|-------------|-------------------------|------------------------|
| $\hat{y}_{1t}$ | Constant    | 0.09796     | 0.07012                 |                        |
|      | $y_{1,t-1}$ | -0.42626    | 0.00874*                | 3.941                  |
|      | $y_{2,t-1}$ | 0.04419     | 0.72512                 | (0.03022)*             |
| $\hat{y}_{2t}$ | Constant    | 0.153619    | 0.0397*                 |                        |
|      | $y_{1,t-1}$ | 0.005699    | 0.9783                  | 3.671                  |
|      | $y_{2,t-1}$ | -0.45619    | 0.0120*                 | (0.03748)*             |

*) 5% significance level

The diagnostics test of VARI(1,1) residuals is done by Jacque-Berra test. The statistics test was 2.8163 with 0.589 of p-value which was greater than the significance level of $\alpha = 5\%$. This means that the distribution of residuals is following multivariate normal distributions. The Portmanteau statistics test was 32.776 with 0.6227 of p-value which was greater than the significance level of $\alpha = 5\%$. This indicated the accomplishment of white noise assumption. The ARCH-Lagrange Multiplier (LM) statistics test was 69 with 0.9511 of p-value which was greater than the significance level of $\alpha = 5\%$.  


This means that residuals were homogenous. It can be concluded that VARI(1,1) model was satisfied residuals assumption.

The model equation of confirmed covid-19 cases Indonesia can be written as follows:
\[
\hat{y}_{1t} = 0.098 - 0.426 y_{1,t-1} + 0.044 y_{2,t-1}.
\]
\[
\hat{z}_{1t} = 1.103 + 1.775 z_{1,t-1} + 1.531 z_{2,t-2} + 1.045 z_{2,t-1} + 0.957 z_{2,t-2}.
\]

The model equation of confirmed covid-19 cases Singapore can be written as follows:
\[
\hat{y}_{2t} = 0.154 + 0.006 y_{1,t-1} - 0.456 y_{2,t-1}.
\]
\[
\hat{z}_{2t} = 1.166 + 1.006 z_{1,t-1} + 0.994 z_{1,t-2} + 1.723 z_{2,t-1} + 1.578 z_{2,t-2}.
\]

The granger causality test is done to investigate causality between confirmed cases in Indonesia and Singapore. The results show that the p-values were 0.7251 and 0.9783 which were greater than the significance level of \( \alpha = 5\% \). This indicated that there is no reciprocal relationship between the confirmed covid-19 cases in Indonesia and Singapore. However, the coefficient correlation between both countries shows that there was a quite strong positive relationship.

According to the VARI(1,1) model, it can be used to obtain the future value of each country. Both countries were forecasted for the next 6 days periods as seen in Table 4. The accuracy can be computed by compared actual data and forecast data which is represented by Mean Absolute Percentage Error (MAPE) value. The MAPE of Indonesia and Singapore series was 32.74 percent and 37 percent, respectively. The accuracy shows that the model is good enough for forecasting covid-19 cases in both countries.

### Table 4. Actual data and forecasting result with VARI(1,1) model.

| Period    | Indonesia | Singapore |
|-----------|-----------|-----------|
|           | Actual    | Forecasting | Actual    | Forecasting |
| 20 Apr '20 | 185       | 352.5544   | 1426     | 856.6403    |
| 21 Apr '20 | 375       | 382.6470   | 1111     | 847.2901    |
| 22 Apr '20 | 283       | 407.3763   | 1016     | 993.8032    |
| 23 Apr '20 | 357       | 440.5734   | 1037     | 1078.3471   |
| 24 Apr '20 | 436       | 471.6781   | 897      | 1212.4815   |
| 25 Apr '20 | 396       | 507.9554   | 618      | 1341.2702   |

### 4. Conclusion

Based on the results and discussion, it can be concluded that the optimum lag was an autoregressive order 1. Therefore, the best model confirmed covid-19 cases of Indonesia and Singapore was VARI(1,1) which significant simultaneously. The covid-19 cases in Indonesia and Singapore were only affected by the previous covid-19 cases in each country. The diagnostics test shows that the VARI(1,1) model has satisfied the multivariate normal, white noise, and homogenous assumption. Meanwhile, the accuracy shows that the model is good enough for forecasting covid-19 cases in both countries.

### 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] Susilo A, Rumende C M, Pitoyo C W, Santoso W D, Yulianti M, Herikurniawan, Sinto R, Singh G, Nainggolan L, Nelwan E J, Chen L K, Widhani A, Wijaya E, Wicaksana B, Maksum M, Annisa F, Jasirwan C OM, Yunihastuti E 2020 Coronavirus Disease 2019: Tinjauan Literatur Terkini Jurnal Penyakit Dalam Indonesia 71
[2] Liu Y, Kuo R, Shih S 2020 COVID-19: The first documented coronavirus pandemic in history *Biomedical Journal*

[3] United Nations 2020 Brief#2: Putting The UN Framework For Socio-Economic Response To COVID-19 Into Action: Insights *United Nations Development Programme (UNDP) Brief Report*

[4] International Labour Organization 2020 A policy framework for tackling the economic and social impact of the COVID-19 crisis *International Labour Organization Policy Brief Report*

[5] Aswicahyono H 2020 Pertumbuhan Eksponensial dan Mitigasi COVID-19: Hindari Framing Dikotomis *CSIS Commentaries DMRU-003*

[6] Kang D, Choi H, Kim J, Choi J 2020 Spatial Epidemic Dynamics of the COVID-19 Outbreak in China *International Journal of Infectious Disease 94* 96-102

[7] Ruchjana B N 2004 Model Vektor Autoregresi dan Penerapannya dalam Prakiraan Produksi Minyak Bumi *Seminar Nasional Matematika Universitas Pendidikan Indonesia*

[8] Hyndman R J and Athanasopoulos G 2018 *Forecasting: principles and practice 2nd edition* (Melbourne: OTexts)

[9] Lutkepohl H 2005 *New Introduction to Multiple Time Series Analysis* (Berlin: Springer Science+Business Media, Inc)

[10] Wei W W S 2006 *Time Series Analysis Univariate and Multivariate Methods Second Edition* (United States: Pearson Addison Wesley)

[11] Tsay R S 2014 *Multivariate Time Series Analysis with R and Financial Application* (New Jersey: John Wiley & Sons)

[12] Hurvich C M and Tsai C 1993 A Corrected Akaike Information Criterion for Vector Autoregressive Model Selection *Journal of Time Series Analysis 14* 3

[13] Wulandari D A, Gusriani N, Ruchjana B N 2014 Penaksiran Parameter Model Vector Autoregressive Integrated (VARI) dengan Metode MLE dan Penerapannya pada Data Indeks Harga Konsumen *IndoMS Journal on Statistics 2* 27-37