The application of vector autoregressive integrated with exogenous variable to model the relationship between covid-19 positive numbers with population growth rate

M F Akbar, E Octaviany, T Toharudin, and B N Ruchjana

1 Department of Statistics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia
2 Badan Pusat Statistik Provinsi Sulawesi Tengah, Jl. Prof Moh Yamin No. 48, Palu 94114, Indonesia
3 Badan Pusat Statistik Kabupaten Sambas, Jl. Pembangunan, Sambas, Indonesia
4 Departement of Mathematics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

Abstract. The Vector Autoregressive Integrated with Exogenous Variable (VARIX) model can be used to examine the relationship between covid-19 positive case numbers (as endogenous variables) and population growth rate (as exogenous variables) while predicting covid-19 positive case numbers. In this paper, we apply the VARIX model to covid-19 phenomena in two regencies/cities in West Java Province, Bandung and Purwakarta. The results of the analysis show that the VARIX model is a suitable model used to describe the intended relationship. The VARIX model has a good forecasting ability characterized by a relatively small MAPE value. The exogenous variables show positive results, so it can be concluded that the positive covid-19 numbers in Bandung City and Purwakarta Regency have a positive correlation with the population growth rate in West Java Province.

1. Introduction
Corona Virus Disease 2019 (covid-19) is becoming a serious challenge in the world of global health including Indonesia. Data from the Task Force for the Acceleration of Handling covid-19 (national data) shows that by the end of April 2020 the number of positive cases of covid-19 in Indonesia reached 10,118 cases, with a total of 792 deaths [1]. The spread of covid-19 cases in Indonesia is becoming more widespread at any time and continues to show a positive trend. Many factors determine the spread of covid-19 cases, one of them is the rate of population growth. According to [2] the birth and death rates not only affect changes in population, but both also the epidemic of diseases. In areas with high population density, this will increase the case of the disease. West Java Province is the province with the highest population in Indonesia, which has more than 49 million people with a population growth rate of 1.36 percent. Meanwhile covid-19 cases in West Java at the end of April numbered 728 cases and the number of deaths reached 92. Even though the population is large, the case of covid-19 in West Java is relatively low. However, strategic actions to accelerate the resolution of covid-19 cases must still be carried out, as well as being a model for other provinces in handling covid-19 cases. Forecasting the number of future covid-19 cases is felt to be significantly needed in order to map appropriate policies so that this pandemic can be completely resolved. Covid-19 data contains trends that form a time series model, one of the multivariate time series data models that can
be used is the VARIX model [3]. The VARIX model is a development of the VARX (Vector Autoregressive with Exogenous Variable) model for non-stationary data by including exogenous variables in the equation [4]. This study will estimate the VARIX model estimation and its application to model the relationship between covid-19 positive numbers and population growth rate, case studies on covid-19 positive numbers in the City of Bandung and Purwakarta Regency.

2. Method
2.1. Data Sources and Research Variables
This study uses data of covid-19 daily positive numbers in the city of Bandung and Purwakarta in the period 3 - 29 April 2019 obtained from the page https://pikobar.jabarprov.go.id. There are some data that are not included, so that the interpolation is done by averaging the positive covid-19 number on the previous day and afterwards. The exogenous variable used in this study is the projected rate of population growth in West Java in 2020 sourced from the Statistics Jawa Barat website https://jabar.bps.go.id. The variables used in this study are explained in the Table 1.

Table 1. Research variables

| Symbol | Information |
|--------|-------------|
| \( Z_{1,t} \) | Positive covid-19 number in Bandung City in period \( t \) |
| \( Z_{2,t} \) | covid-19 positive numbers in Purwakarta Regency in the period \( t \) |
| \( X_t \) | Projected rate of population growth in West Java in 2020 |
| \( Y_{1,t} \) | Positive covid-19 number in Bandung City in the \( t \) period after the process of differencing was carried out |
| \( Y_{2,t} \) | A positive number of covid-19 in Purwakarta Regency in the \( t \) period after differencing process was conducted |

2.2. VARIX order 1
The VARIX model is a development of the VARX model with non-stationary data. The form of the VARIX model (1,1) is stated as follows:

\[
\bar{Y}_t = \phi_{(1)}\bar{Y}_{t-1} + \phi_{(2)}\bar{Y}_{t-2} + \ldots + \phi_{(p)}\bar{Y}_{t-p} + \beta_1\bar{X}_t + \ldots + \beta_q\bar{X}_{t-q} + \bar{\epsilon}_t, \tag{1}
\]

Estimating the parameters of the VARIX model (1,1) uses the Ordinary Least Squares OLS method can be described as follows:

1. Take the equation \( \bar{Y} = X\bar{\phi} + \bar{\epsilon} \), \( \bar{\epsilon} \sim N(0, \sigma^2) \)
2. Multiply the two segments with \( X^T \) so that \( (X^T)\bar{Y} = (X^T)X\bar{\phi} \) is obtained
3. Multiply the two segments with \( (X^TX)^{-1} \) so \( (X^TX)^{-1}(X^T)\bar{Y} = (X^TX)^{-1}(X^T)X\bar{\phi} \) is obtained

So that \( \bar{\phi} = (X^TX)^{-1}(X^T)\bar{Y} \), where:

\[
\bar{\phi} = \begin{bmatrix}
\phi_{11} \\
\phi_{12} \\
\phi_{13} \\
\phi_{14} \\
\beta_1 \\
\beta_2
\end{bmatrix}, \quad X = \begin{bmatrix}
Y_{1,2} & Y_{2,2} & 0 & 0 & 0 & 0 \\
0 & 0 & Y_{1,2} & Y_{2,2} & 0 & 0 \\
Y_{1,3} & Y_{2,3} & 0 & 0 & 0 & 0 \\
0 & 0 & Y_{1,3} & Y_{2,3} & 0 & 0 \\
0 & 0 & 0 & 0 & X_{1,3} & 0 \\
0 & 0 & 0 & 0 & 0 & X_{1,4}
\end{bmatrix}, \quad \bar{Y} = \begin{bmatrix}
Y_{1,3} \\
Y_{2,3} \\
Y_{1,4} \\
Y_{2,4} \\
Y_{1,3} \\
Y_{1,4}
\end{bmatrix}
\]

Then estimate the parameters of the VARIX model (1,1). Through Maple software, it is obtained:

\[
\begin{bmatrix}
\phi_{11} \\
\phi_{12} \\
\phi_{13} \\
\phi_{14} \\
\beta_1 \\
\beta_2
\end{bmatrix} = \begin{bmatrix}
Y_{1,2}^2 + Y_{2,2}^2 & 0 & Y_{1,2}Y_{1,3} + Y_{2,2}Y_{2,3} & 0 & 0 & 0 \\
Y_{1,2}^2 + Y_{2,2}^2 & 0 & Y_{1,2}Y_{1,3} + Y_{2,2}Y_{2,3} & 0 & 0 & 0 \\
Y_{1,3}^2 + Y_{2,3}^2 & 0 & Y_{1,3}Y_{1,4} + Y_{2,3}Y_{2,4} & 0 & 0 & 0 \\
Y_{1,3}^2 + Y_{2,3}^2 & 0 & Y_{1,3}Y_{1,4} + Y_{2,3}Y_{2,4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}^{-1}
\]
\[
\begin{pmatrix}
Y_{1,2}Y_{1,3} + Y_{1,3}Y_{1,4} \\
Y_{1,3}Y_{2,2} + Y_{2,3}Y_{1,4} \\
Y_{1,2}Y_{2,3} + Y_{1,3}Y_{2,4} \\
Y_{2,2}Y_{2,3} + Y_{2,3}Y_{2,4} \\
X_{1,3}Y_{1,3} \\
X_{1,4}Y_{1,4}
\end{pmatrix}
\]

Then obtained the VARIX model parameters (1,1) as follows:
\[
\hat{\phi}_{11} = \frac{(Y_{1,3}^2 + Y_{2,3}^2)(Y_{1,2}Y_{1,3} + Y_{1,3}Y_{1,4}) - (Y_{1,2}Y_{1,3} + Y_{2,2}Y_{2,3})(Y_{1,2}Y_{2,3} + Y_{1,3}Y_{2,4})}{Y_{1,2}^2Y_{2,3}^2 - 2Y_{1,2}Y_{1,3}Y_{2,2}Y_{2,3} + Y_{2,2}^2Y_{1,3}^2}
\]
\[
\hat{\phi}_{12} = \frac{(Y_{1,3}^2 + Y_{2,3}^2)(Y_{2,2}Y_{1,3} + Y_{2,3}Y_{1,4}) - (Y_{1,2}Y_{1,3} + Y_{2,2}Y_{2,3})(Y_{2,2}Y_{2,3} + Y_{2,3}Y_{2,4})}{Y_{1,2}^2Y_{2,3}^2 - 2Y_{1,2}Y_{1,3}Y_{2,2}Y_{2,3} + Y_{2,2}^2Y_{1,3}^2}
\]
\[
\hat{\phi}_{21} = \frac{(Y_{1,2}^2 + Y_{2,2}^2)(Y_{1,2}Y_{2,3} + Y_{1,3}Y_{1,4}) - (Y_{1,2}Y_{1,3} + Y_{2,2}Y_{2,3})(Y_{1,2}Y_{2,3} + Y_{1,3}Y_{1,4})}{Y_{1,2}^2Y_{2,3}^2 - 2Y_{1,2}Y_{1,3}Y_{2,2}Y_{2,3} + Y_{2,2}^2Y_{1,3}^2}
\]
\[
\hat{\phi}_{22} = \frac{(Y_{1,3}^2 + Y_{2,3}^2)(Y_{1,2}Y_{1,3} + Y_{2,3}Y_{2,4}) - (Y_{1,2}Y_{1,3} + Y_{2,2}Y_{2,3})(Y_{2,2}Y_{1,3} + Y_{2,3}Y_{1,4})}{Y_{1,2}^2Y_{2,3}^2 - 2Y_{1,2}Y_{1,3}Y_{2,2}Y_{2,3} + Y_{2,2}^2Y_{1,3}^2}
\]
\[
\hat{\beta}_1 = \frac{Y_{1,3}}{X_{1,3}}
\]
\[
\hat{\beta}_2 = \frac{Y_{1,4}}{X_{1,4}}
\]

2.3. Stages of Data Analysis
The steps of analysis in this study are as follows [5]:
1. Examine a general description of the data by plotting daily covid-19 figures in the City of Bandung and Purwakarta Regency.
2. Perform univariate stationary objectives for each endogenous variable.
3. Perform differencing processes for each endogenous variable.
4. After the Autoregressive Integrated (ARI) model \((p, d)\) is obtained, check the correlation between endogenous variables. If there is a strong correlation, then the addition of exogenous variables can be done to get the VARIX model \((p, d)\)
5. Estimate the parameters
6. Do the Diagnostic Checking on the model
7. Do MAPE forecasting and calculating

3. Result and Discussion
3.1. Data Stationarity Test
The time series data plot using R 3.6.3 Software can be seen in Figure 1.

Figure 1. Time series, acf, and pacf positive plot data covid-19 in the city of Bandung
It can be seen generally that the positive number of covid-19 in Bandung has not been stationary. However, to ensure the stationarity of the data, an Augmented Dickey-Fuller (adf) test is then performed. The adf test results showed a $p$-value of 0.2105 (greater than $\alpha = 0.05$), so it can be concluded that the data tested was not stationary.

**Figure 2.** Time series, acf, and pacf positive plot data covid-19 in Purwakarta

Similar as it is in the City of Bandung, it can be generally seen that the positive number of covid-19 in Purwakarta Regency is not stationary (Figure 2). However, to ensure the stationarity of the data, an Augmented Dickey-Fuller (adf) test is then performed. The adf test results showed a $p$-value of 0.1738 (greater than $\alpha = 0.05$), so it can be concluded that the data tested was not stationary.

### 3.2. Differencing Process

The differencing process is carried out on each endogenous variable. The time series data plot of differencing results using R 3.6.3 software can be seen in Figure 3.

**Figure 3.** The plot of differencing results, ACF, and PACF covid-19 positive numbers in the City of Bandung

**Figure 4.** The Plot of differencing results, ACF, and PACF covid-19 positive numbers in Purwakarta Regency
From the Figure 4, it can be seen that the pacf differencing results in Bandung City and Purwakarta Regency data are both cut off in the 3rd lag. To form a Vector Autoregressive Integrated (VARI) model, a strong correlation between endogenous variables with the same model is required [6]. Inferential analysis was performed to see the stationarity of the differencing results using the ADF test. By using a significance level of 5 percent, the following results are obtained in Table 2.

Table 2. ADF test results for covid-19 positive numbers in Bandung City and Purwakarta Regency

| Variables | P-value | Information  |
|-----------|---------|--------------|
| \(Y_{1,t}\) | 0.01    | stationary   |
| \(Y_{2,t}\) | 0.01    | stationary   |

Table 3. Correlation between endogenous variables

| Variables | \(Z_{1,t}\) | \(Z_{2,t}\) |
|-----------|------------|------------|
| \(Z_{1,t}\) | 1          | 0.894      |
| \(Z_{2,t}\) | 0.894      | 1          |

Based on the Table 2 above it appears that there is a strong relationship between the two endogenous variables, so that VARI models can be formed. Akaike introduced the best model selection criteria by considering the number of parameters in the model [7], namely the Akaike Information Criterion (AIC), this selection criterion is based on the smallest (minimum) AIC value among the existing models. AIC can be formulated as follows:

\[
AIC(M) = n \ln(\hat{\sigma}^2_p) + 2M
\]  

where \(M\) is the number of parameters in the model. The optimal order in the model is chosen based on the value of \(M\), which is a function of \(p\) and \(q\), so that the value of \(AIC(M)\) is the minimum. Based on the R output, the \(p\)-value of the AIC value is significant (less than 0.05) at the 3rd lag (0.0454), so that the VARI model (3.1) is suitable for use in research.

3.3. VARIX Model Parameter Estimation (3.1)

Using the OLS method, the estimated parameter coefficient for the VARIX model (3.1) yields the estimated parameters as represented in Table 3.

Table 3. VARIX model parameter estimation results (3.1) using software R 3.6.3

| Parameter | Estimate |
|-----------|----------|
| \(\hat{\phi}_{11}(1)\) | -0.47    |
| \(\hat{\phi}_{12}(1)\) | 2.69     |
| \(\hat{\phi}_{21}(1)\) | -0.05    |
| \(\hat{\phi}_{22}(1)\) | 0.35     |
| \(\hat{\phi}_{11}(2)\) | -0.52    |
| \(\hat{\phi}_{12}(2)\) | 3.84     |
| \(\hat{\phi}_{21}(2)\) | -0.03    |
| \(\hat{\phi}_{22}(2)\) | 0.22     |
| \(\hat{\phi}_{11}(3)\) | -0.55    |
| \(\hat{\phi}_{12}(3)\) | -3.26    |
| \(\hat{\phi}_{21}(3)\) | 0.02     |
| \(\hat{\phi}_{22}(3)\) | -0.62    |
So the equation becomes:
\[
\hat{Y}_{1,t} = -0.47Y_{1,t-1} + 2.69Y_{2,t-1} - 0.52Y_{1,t-2} + 3.84Y_{2,t-2} - 0.55Y_{1,t-3} - 3.26Y_{2,t-3}
\]
\[
\hat{Y}_{2,t} = -0.05Y_{1,t-1} + 0.35Y_{2,t-1} - 0.03Y_{1,t-2} + 0.22Y_{2,t-2} - 0.02Y_{1,t-3} - 0.62Y_{2,t-3}
\]
When \( \hat{Y}_t = \hat{Z}_t - Z_{t-1} \), then the results of the equation for the positive covid-19 numbers in Bandung City and Purwakarta Regency are as follows:
\[
\hat{Z}_{1,t} = 0.53Z_{1,t-1} + 2.69Z_{2,t-1} - 0.05Z_{1,t-2} + 1.15Z_{2,t-2} - 0.03Z_{1,t-3} - 7.1Z_{2,t-3} + 0.55Z_{1,t-4} +
\]
\[
3.26Z_{2,t-4} + 10.01X_{1,t}
\]
\[
\hat{Z}_{2,t} = -0.05Z_{1,t-1} + 1.35Z_{2,t-1} + 0.22Z_{1,t-2} - 0.13Z_{2,t-2} + 0.01Z_{1,t-3} - 0.84Z_{2,t-3} +
\]
\[
0.02Z_{1,t-4} + 0.62Z_{2,t-4} + 0.58X_{1,t}
\]

Testing the model simultaneously using the F test shows that the \( p \)-value is 0.0042 and 0.0649 smaller than the significance level of 10% which means that both equations are simultaneously significant. The two equations above show that the observed variables, namely the positive covid-19 number in Bandung, the positive covid-19 number in Purwakarta Regency, and the rate of population growth in West Java influence each other.

### 3.4. Diagnostic Checking

#### 3.4.1. Multivariate Normal Distributed Residual Test

Multivariate normally distributed residual assumption test is performed to see whether the residuals of the multivariate normally distributed model use the Multivariate Jarque-Berra (JB) test [8]. The results of the Multivariate JB test on the VARI model (3, 1) are as follows:

**Table 4.** The result of multivariate JB test

| Test     | Chi-square | df | P-value |
|----------|------------|----|---------|
| JB       | 1,1582     | 4  | 0.8849  |
| Skewness | 0.3092     | 2  | 0.8567  |
| Kurtosis | 0.8490     | 2  | 0.6541  |

Table 4 shows that the results of the Multivariate JB test shows \( p \)-value = 0.8849 greater than the significance level of 10%. It can be concluded that \( H_0 \) is rejected, which means that the assumption of multivariate normal distribution in the VARI model residuals is fulfilled.

#### 3.4.2. Multivariate White Noise Residual Test

Multivariate white noise residual assumption test is performed to see whether the residuals of the model are no longer correlated with each other using the Portmanteau test [9]. The Portmanteau test results on the VARI model (3, 1) are as follows:

**Table 5.** The result of Portmanteau test

| Chi-square | df | P-value |
|------------|----|---------|
| 18,288     | 28 | 0.9187  |

The results of the assumption of multivariate white noise residuals in Table 5 show that the \( p \)-value = 0.6321 is greater than the significance level of 10%. It can be concluded that \( H_0 \) is accepted, which means that the assumption of multivariate white noise in residual VARI models has been fulfilled.
3.4.3. Residual Homogeneity Test

Test the assumption of residual homogeneity using the ARCH-Lagrange Multiplier (LM) test [10]. The ARCH-LM test results on the VARI model (3,1) are as follows:

| Table 6. The result of ARCH-LM test |
|------------------------------------|
| Chi-square | df | P-value |
| 33         | 90 | 1       |

The results of the residual homogeneity assumption test in Table 6 show that $p$-value = 1 is greater than the 10% significance level, so it can be concluded that $H_0$ is accepted, which means that the assumption of homogeneity in the VARI model residuals has been fulfilled.

3.4.4. MAPE Forecasting and Calculation

The next stage of analysis in VARIX modeling is calculating the comparison of accuracy in forecasting, where the method chosen to be used in comparing the measurement of accuracy with Mean Absolute Percentage Error (MAPE). MAPE values are used to explain the accuracy of the estimates in the VARIX model [11]. Comparison of the accuracy of the estimates through the results of the calculation of the MAPE value of the VARIX model (3,1) at the appropriate time lag can be shown in Figure 5.

![Actual data and forecast positive covid-19 cases in Bandung City](image1)

![Actual data and forecast positive covid-19 cases in Purwakarta Regency](image2)

**Figure 5. MAPE current data and covid-19 prediction in Bandung and Purwakarta Regency**

According to the calculations, the MAPE value for Bandung City is 1%, while for Purwakarta Regency is 11.8%. Both values tend to be small so that it can be said that the VARIX model (3,1) is feasible to be applied to the research data.

4. Conclusion

The VARIX model obtained to illustrate the relationship between the positive covid-19 number in Bandung City and Purwakarta Regency with the population growth rate in West Java Province is VARIX (3,1). Based on the MAPE values obtained, it can be concluded that the VARIX model (3,1) has a good forecasting ability so that it can be used in research data. Exogenous variables show a positive value, meaning that there is a positive correlation between the covid-19 number and the rate of population growth.

Acknowledgments

The authors thank Rector Universitas Padjadjaran who provided financial assistance for the dissemination of research through Academic Leadership Grant 2020 with contract number: 1427/UN6.3.1/LT/2020. Also thanks to the reviewers for the valuable review of this paper.
References

[1] Kementerian Kesehatan Republik Indonesia 2020 Situasi terkini perkembangan coronavirus disease (COVID-19) 30 April 2020 Online: https://covid19.kemkes.go.id/situasi-infeksi-emerging/info-corona-virus/situasi-terkini-perkembangan-coronavirus-disease-covid-19-30-april-2020

[2] Djallalluddin, Hasni H B, Riana W, and Lisda H 2004 Gambaran penderita kejadian luar biasa demam berdarah dengue di kabupaten Banjar dan kota Banjarbaru tahun 2001 DEXA MEDIA 17(2) 591

[3] Hamilton J D 1994 Time series analysis (New Jersey: Princeton University Press)

[4] Abraham B and Ledolter J 1983 Statistical methods for forecasting (New Jersey: John Wiley and Sons Inc)

[5] Makridakis S, Wheelwright, Steven C, and McGee V E 1978 Forecasting: Methods and Applications (Canada: John Wiley and Sons Inc)

[6] Hendarsin O P and Hidayat M 2019 Penerapan model vector autoregressive integrated with exogenous untuk memodelkan hubungan antara jumlah penduduk miskin dengan angka harapan hidup Bandung: Paper Program Studi Magister Statistika Universitas Padjadjaran.

[7] Dewi C N 2017 Penerapan model vector autoregressive integrated with exogenous pada data debit air tiga lokasi sungai di Jawa Barat Jatinangor: Skripsi Program Studi Matematika, Universitas Padjadjaran.

[8] Wei and William W S 2006 Time series analysis: Univariate and Multivariate Methods (New York: Pearson Addison Wesley)

[9] Lutkepohl H 2005 New introduction to multiple time series analysis (Berlin: Springer-Verlag)

[10] Cryer, Jonathan D, and Kung S C 2008 Time series analysis with applications in R Second Edition (New York: Springer Science+Business Media, LLC)

[11] Tsay and Ruey S 2014 Multivariate time series analysis: with R and financial applications (New Jersey: John Wiley & Sons Inc)