BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2022 Volume 16 Issue 2 Page 751–760
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss2pp751–760

COMPARISON OF AUTOREGRESSIVE MODEL WITH MISSING DATA TREATED USING ORDINARY LEAST SQUARES AND INTERPOLATION WITH WEIGHTING METHOD

Syifani Akmaliah [1], Dianne Amor Kusuma[2], Budi Nurani Ruchjana[3*]

[1,2,3]Mathematics Department, Faculty of Mathematics and Natural Sciences,Universitas Padjadjaran Bandung-Sumedang St., Jatinangor, Sumedang, 45363, Indonesia

Corresponding author's e-mail: [3] budi.nurani@unpad.ac.id*

Abstract. *Bandung is committed to contributing to the achievement of the Sustainable Development Goals (SDGs) in Indonesia. One of the efforts that can be made to support the 13th pillar of SDGS regarding climate change is to forecast the air temperature of Bandung City in the future. One of the models that can be used for forecasting air temperature data in Bandung is the Autoregressive (AR) model. Based on BMKG data, often the time series data obtained has missing data. Therefore, in order to do a good time series analysis, it is necessary to make an effort to correct the missing data. The purpose of this research was to examine the procedure for overcoming missing data in the AR model using the Ordinary Least Squares (OLS) method and Interpolation with Weighting, which was applied to forecasting the average air temperature data in the city of Bandung. The research methodology followed the Box-Jenkins 3-step procedure. The first-order AR estimation parameter model was estimated using the OLS method and then used to overcome missing data using both methods with weighting using R software. Both methods resulted in an estimated value of 0.9991 and the same Mean Average Percentage Error (MAPE) value of 2,459% with very accurate criteria. Therefore, to overcome the missing data on the average air temperature data in the city of Bandung with a parameter estimator close to one, we got the same result for both methods.*

Keywords: *autoregressive, average air temperature data, Interpolation method with weighting, missing data, OLS method.*

Article info:

Submitted: 28th February 2022 *Accepted: 20th May 2022*

How to cite this article:
S. Akmaliah, D. A. Kusuma, and B. N Ruchjana, "COMPARISON OF AUTOREGRESSIVE MODEL WITH MISSING DATA TREATED USING ORDINARY LEAST SQUARES AND INTEPOLATION WITH WEIGHTING METHOD", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 2, pp. 751-760, June, 2022.

1. INTRODUCTION

The Sustainable Development Goals (SDGs) are a series of sustainable development agendas that were inaugurated as a new universal development agenda on September 25th, 2015 by UN member countries, including Indonesia. One of the targets of the 13th pillar of SDGs is climate change planning and management [1]. Climate change is a change that occurs significantly in climate, air temperature, and rainfall ranging from decades to millions of years [2].

The city of Bandung is committed to contributing to the SDGs in Indonesia [3]. One of the efforts that can be made to support the 13th pillar of SDGs is to forecast the air temperature of Bandung City in the future using past data as a reference for the government in planning and managing climate change. Time series analysis for univariate data can be done using the Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) model. One of the Box-Jenkins models for stationary data is the AR model. The AR model is a Box-Jenkins time series model that is often used in time-series data analysis. In the AR model, the value of Z at time t is expressed as a linear combination of the values at the previous time plus the error factor [4]. Based on BMKG data, often the time series of data obtained has missing data. Lost data can be caused by information on an object that is not provided, if the information is difficult to find or does not exist. So, in order to perform the Box-Jenkins method properly, efforts need to be made to identify and correct missing data.

Research on the AR model has been carried out previously, including [5] developing the ARIMA model. Including [6] examines missing values in time series which are overcome using Polynomial Curve Fitting, Cubic Spline, ARIMA interpolation, and Structural Time Series Analysis methods. In [7] examined the AR model with missing data, which was overcome using the Ordinary Least Squares method and its application to the rupiah exchange rate data. In [8] compared the AR model with missing data that was overcome using the Yule-Walker Estimator, Kalman Filter, Online Gradient Descent, Attribute Efficient Ridge Regression, and Autoregressive Least Squares Impute method with an experimental study. Based on the explanation above, the writer is interested in studying the differences between the AR model with missing data which is overcome using the Ordinary Least Squares and Interpolation with Weighted methods.

The purpose of this study is to examine the procedure for overcoming missing data in the AR model using the OLS method and Interpolation with Weighting, which is applied to forecasting the average air temperature data in the city of Bandung. According to the World Meteorological Organization, air temperature is one of the parameters of the climate that needs to be observed routinely, measured at weather or climate observation stations spread throughout the world every day at certain times with the instructions of the World Meteorological Agency [9]. In this research, R software was used to model the average air temperature data in the city of Bandung using the AR model with missing data. Software R is a complete statistical data analysis system as a result of the collaborative research of various statisticians around the world. The software R can be found for free at the CRAN-archive, which is the Comprehensive R Archive Network at http://cran.r-project.org [10].

2. RESEARCH METHODS

This research uses data on the daily average air temperature in Bandung City from January 02 to May 22, 2021. The data was obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG). The software used in this research is R Studio using library time series.

2.1 Procedure to Handling Missing Data using Ordinary Least Squares And Intepolation With Weighting Method

1. Input data.
2. Plot time series.
3. Checking missing data with Software R.
4. Setting aside time for observation of missing data, namely the *i*-th data.
5. Checking stationary.

Stationarity is a set of conditions that allow estimation of model parameters whose properties are standard, for example t test statistics that have approximately normal distributions in large samples. The data are not hugging that trend line tightly enough to be considered stationary [11].

a. Using Augmented DickeyFuller (ADF), check the stationary of the data in the mean.

The augmented Dickey-Fuller (ADF) test is a formal statistical test done to ensure stationary. The Augmented Dickey-Fuller Test (ADF) is unit root test for stationarity [12].

b. Using the Box-Cox transformation, check the stationary of the data in the variance.

This is the most advanced transformation technique introduced by two statisticians named as George Box and Sir David Cox and known as Box-Cox transformation. George Box and Sir David Cox developed a procedure to identify an appropriate exponent (λ = l) for transforming data into a normal shape [13].

6. Identify the order of AR models with ACF plots and PACF plots.
7. Estimating the AR model parameters using the OLS method.
8. Handling missing data

a. Handling missing data using ordinary least squares method.

Missing data in the AR(1) model can be handled by using the OLS method with the following equations [7] :

$$\widehat{Z_s} = \frac{\varphi_1}{1+\varphi_1^2} Z_{t+1} + \frac{\varphi_1}{1+\varphi_1^2} Z_{t-1} \quad (1)$$

Where:

Z_t : Stationary time series

φ_1 : AR(1) parameter

e_t : Error at time *t*

$\widehat{Z_s}$: Missing data

b. Handling missing data using Interpolation with Weighting method.

Missing data in the AR(1) model can be handled by using the interpolation with weighting method with the following equations [6] :

$$\widehat{h_t} = \frac{1}{1+\varphi_1^2} Z_{t+1} + \frac{1}{1+\varphi_1^2} Z_{t-1} \quad (2)$$

where,

Z_t : Stationary time series

φ_1 : AR(1) parameter

e_t : Error at time *t*

$\widehat{h_t}$: Missing data

9. Obtain the estimated value of the missing data using the OLS and Interpolation with Weighting method.

2.2 Box-Jenkins Procedure for Implementing an AR Model with Missing Data

1. Input the calculated *i*-th data value to fill in the gaps in the original dataset.
2. Checking stationary.
    a. Using Augmented DickeyFuller (ADF), check the stationary of the data in the mean.
    b. Using the Box-Cox transformation, check the stationary of the data in the variance.
3. Identify the order of AR models with ACF plots and PACF plots.
4. Estimating the AR model parameters using the OLS method.

5. Check for normality by looking at the quartiles (Q-Q) plot.
6. Forecasting for the next seven days based on average air temperature data in Bandung using the AR(1) model estimate.
7. Evaluate the time series model.

   To evaluate the time series model, the method Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Error (ME), Root Mean Squared Error (RMSE) can be used [14]. In this study the method used is MAPE. The MAPE equation [4]:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Z_t - \widehat{Z_t}}{Z_t}\right| \quad (3)$$

   with,

   Z_t : Random variable at time t

   t : Time t

   n : Number of time periods
8. Comparing the MAPE value of the AR model with missing data treated using OLS and Interpolation with Weighting method.

3. RESULTS AND DISCUSSION

3.1. Descriptive Statistics of Bandung City's Average Air Temperature Data

In this research, the data has one missing data on the 63rd observation on March 5, 2021. The data has an average value of 23.62°C. The plot of time series data can be seen in the following figure:

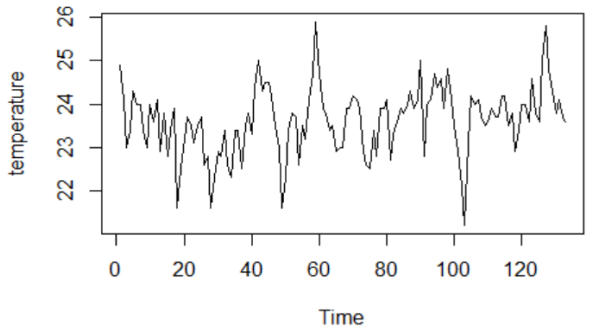

Figure 1. Time Series Plot of the Average Air Temperature Data in the City of Bandung

Figure 1 shows the data plots are interrupted at 60–80 intervals. This shows that there is missing data at intervals of 60–80 whose location is not yet known

3.2. Checking Missing Data

Using The software R, it was discovered that the average air temperature data in the city of Bandung has one missing data, namely the 63rd observation. The distribution graph for missing data is presented in the following figure:

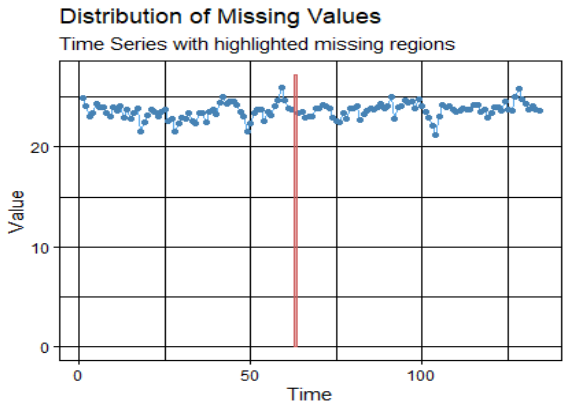

Figure 2. Distribution Graph on Missing Data on Average Air Temperature Data in the City of Bandung

Figure 2 shows that there is a vertical red line that describes the location of the missing data. The missing data lies between the 50-100 observation interval, according to the information that the missing data was found in the 63rd observation.

3.3. Setting aside Missing Data Observation Time

Average air temperature data in the City of Bandung has missing data. Furthermore, the time observation of the missing data, namely the 63rd observation, is set aside so that the Box-Jenkins iteration can be carried out. The initial data was 134 observations to 133 observations.

3.4. Checking Stationary

Stationary data is a condition that must be met in conducting time series analysis. The identification of the stationary data in the mean was carried out using the Augmented Dickey Fuller (ADF) test, which obtained a *P-value* of 0,01 so that the data was said to be stationary in the mean. The identification of the stationary data in the variance was carried out using the Box-Cox transformation obtained, which a value of 1,22677> 1, meaning that the data is stationary in variance.

3.5. Identify the Order of AR Models with ACF Plots and PACF Plots

Identification of the AR(1) model for average air temperature data in the city of Bandung using ACF and PACF plots. The following ACF and PACF plots were created using the software R:

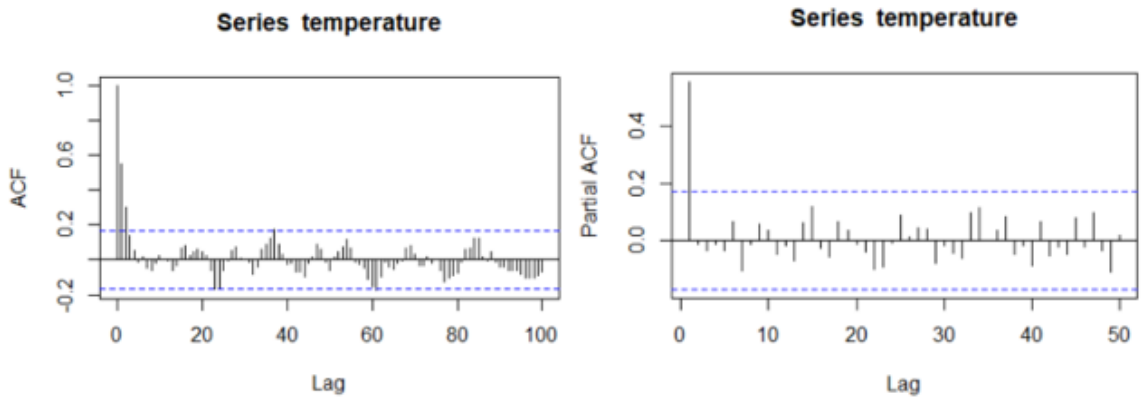

Figure 3. ACF and PACF Plots of Average Air Temperature Data in the City of Bandung with Time Observation of Missing Data have been Set Aside

Figure 3 shows the ACF plot of average air temperature data in the city of Bandung shows tails-off pattern. While the PACF plot forms a cut-off pattern in the 1st lag.

3.6. Estimating the AR Model Parameters using the OLS Method

Estimating the parameters of the AR(1) model using the OLS method with software R produces a parameter value of 0.9991. This shows that the data on the average air temperature of the city of Bandung is influenced by data from one previous observation of 0.9991. The parameter estimate provides an AR(1) model for the average air temperature data of the city of Bandung with missing data that has been set aside in the following equation:

$$\hat{Z}_t = 0.9991\, Z_{t-1} \qquad (3)$$

3.7. Handling Missing Data using Ordinary Least Squares Method

To handle missing data in the AR(1) model using the method using the OLS method, the following equation (1) is used:

$$\widehat{Z_s} = \frac{\varphi}{1+\varphi^2} Z_{t+1} + \frac{\varphi}{1+\varphi^2} Z_{t-1}$$

$$\widehat{Z_s} = \frac{0.9991}{1+0.9991^2}(23{,}4) + \frac{0.9991}{1+0.9991^2}(23.8) = 23.59999$$

The results of the calculation of missing data using software R, namely the 63rd data of 23.59999.

3.8. Handling Missing Data using Interpolation with Weighting Method

To handle missing data in the AR(1) model using the method using the Interpolation with Weighting method, the following equation (2) is used:

$$\widehat{h_t} = \frac{1}{1+\varphi^2} Z_{t+1} + \frac{1}{1+\varphi^2} Z_{t-1}$$

$$\widehat{h_t} = \frac{1}{1+0.9991^2}(23.4) + \frac{1}{1+0.9991^2}(23.8) = 23.62125$$

The results of the calculation of missing data using software R, namely the 63rd data of 23.62125. Therefore, the difference in the estimated value of missing data using OLS and Interpolation with Weighting is 0.02126.

The estimated value of the 63rd observation data was calculated using OLS and Interpolation with weighting method inputted to the initial data. Subsequent time series analysis followed the Box-Jenkins process for univariate data.

3.9. Box-Jenkins Procedure for Implementing AR(1) Model for Data with Missing Data has been Handled using the OLS Method

The Box-Jenkins AR(1) model procedure for data with missing data has been handled using the OLS method carried out with software R. The first step is to test the stationary of the data in the mean and variance. For 7 iterations, the ADF test p-value is 0.01 and the Box-Cox transformation value is 1.999924, so that for each iteration the data is stationary in the mean and variance.

The next step is to identify the AR(1) model with ACF and PACF plots. For 7 iterations the resulting ACF pattern tails off and PACF describes the cut off at lag-1 can be identified as the AR(1) model. The parameter estimation results obtained a value of 0.9991 in the range -1 to 1 can be said to be stationary. Next, a diagnostic error check is carried out by checking the normality of the Q-Q Plot. The results of the error normality check are presented as follows:

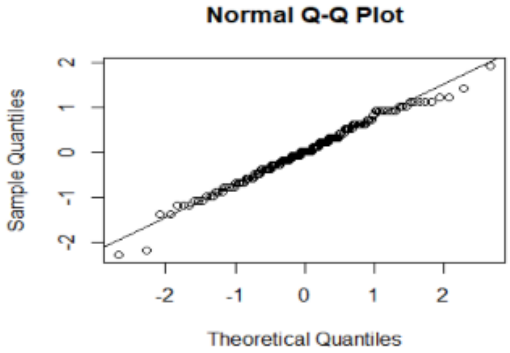

Figure 4. Q-Q Plot Average Air Temperature Data in the City of Bandung with Missing Data that has been Handled Using OLS 1st Iteration

Figure 4 shows that average air temperature data in the City of Bandung with missing data has been handled using OLS method 1st iteration is normal because it lies on the normal line and only a few points are located outside the normal line.

The next step is forecast. Forecasting is done for t = 136, ...,141. The results are presented in Table 1.

Table 1. Forecasting Results 1 to 7 of the OLS Method

	Forecasting Value
1st Iteration	$\hat{Z}_{135}$ =23.6
2nd Iteration	$\hat{Z}_{136}$ =23.6
3rd Iteration	$\hat{Z}_{137}$ =23.5
4th Iteration	$\hat{Z}_{138}$ =23.5
5th Iteration	$\hat{Z}_{139}$ =23.5
6th Iteration	$\hat{Z}_{140}$ =23.5
7th Iteration	$\hat{Z}_{141}$ =23.5

Forecasting and actual plot data are presented in Figure 5.

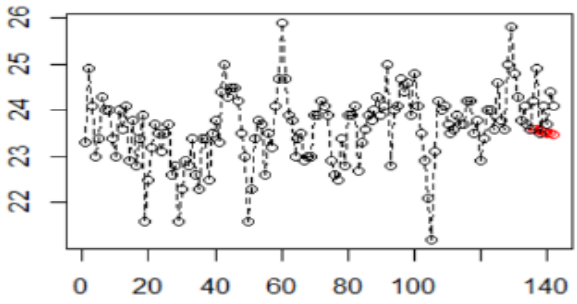

Figure 5. Plot of Forecasting Results and Actual Data for the OLS Method

Figure 5 shows the black line is the actual data and the red dots is the forecasting results, the forecasting results that are formed are close to the actual data points. The resulting MAPE value of 2.549% is categorized as very accurate. This shows that the model forecasted the data very accurately.

3.10. Box-Jenkins Procedure for Implementing AR(1) Model for Data with Missing Data has been Handled using the Interpolation with Weighting Method

The Box-Jenkins AR(1) model procedure for data with missing data has been handled using the Interpolation with Weighting method carried out with software R. The first step is to test the stationary of the data in the mean and variance. For 7 iterations, the ADF test p-value is 0.01 and the Box-Cox transformation value is 1.999924, so that for each iteration the data is stationary in the mean and variance.

The next step is to identify the AR(1) model with ACF and PACF plots. For 7 iterations the resulting ACF pattern tails off and PACF describes the cut off at lag-1 can be identified as the AR(1) model. The parameter estimation results obtained a value of 0.9991 in the range -1 to 1 can be said to be stationary. Next,

a diagnostic error check is carried out by checking the normality of the Q-Q Plot. The results of the error normality check are presented as follows:

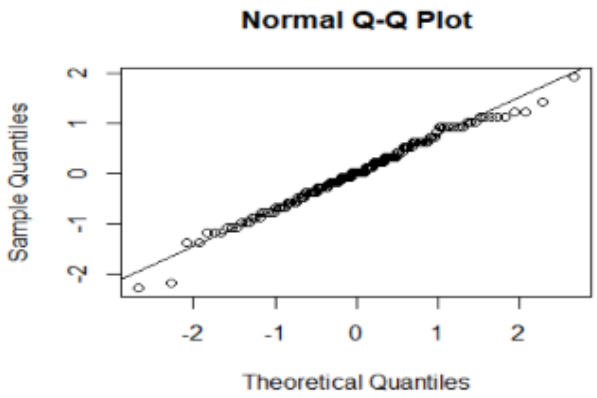

Figure 6. Q-Q Plot Average Air Temperature Data in the City of Bandung with Missing Data that has been Handled Using Interpolation with Weighting 1st Iteration

Figure 6 shows average air temperature data in the City of Bandung with missing data has been handled using Interpolation with Weighting method 1st iteration is normal because it lies on the normal line and only a few points are located outside the normal line.

The next step is forecast. Forecasting is done for t = 136, ...,141. The results are presented in Table 2.

Table 2. Forecasting Results 1 to 7 of the Interpolation with Weighting Method

	Forecasting Value
1st Iteration	$\hat{Z}_{135}$ =23.6
2nd Iteration	$\hat{Z}_{136}$ =23.6
3rd Iteration	$\hat{Z}_{137}$ =23.5
4th Iteration	$\hat{Z}_{138}$ =23.5
5th Iteration	$\hat{Z}_{139}$ =23.5
6th Iteration	$\hat{Z}_{140}$ =23.5
7th Iteration	$\hat{Z}_{141}$ =23.5

Forecasting and actual plot data are presented in Figure 7.

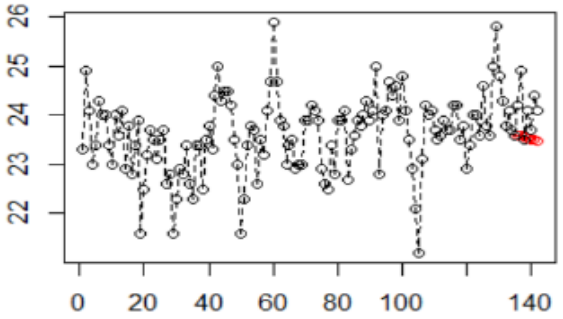

Figure 7. Plot of Forecasting Results and Actual Data for the Interpolation with Weighting Method

Figure 7 shows that the black line is the actual data and the red dots is the forecasting results, the forecasting results that are formed are close to the actual data points. The resulting MAPE value of 2.549% is categorized as very accurate. This shows that the model forecasted the data very accurately.

3.11. Box-Jenkins Procedure for Comparison of Autoregressive Model with Missing Data Treated using Ordinary Least Squares and Interpolation with Weighting Method

The estimated value of the 63rd data using the OLS method has a value of 23.59999, while the estimated value of the 63rd data using the Interpolation method with Weighting has a value of 2.62125, meaning that it has a very small difference of 0.02126. This is because in the AR(1) model the value is close to 1, so the estimated

missing data value obtained using OLS and Interpolation with Weighting has a very small difference. Furthermore, based on the data analysis of the AR model with missing data that is overcome using OLS and Interpolation with Weighting , the MAPE value is the same as the very accurate category.

4. CONCLUSIONS

Based on the results of the discussion obtained, it is concluded that

1. Handling missing data in the AR(1) model using the OLS and Interpolation with Weighting method can be done through the data checking procedure first. Then the data is estimated for the model parameters using OLS. The estimated value is used to determine missing data using equation (1) for OLS and equation (2) for interpolation with weighting.
2. The application of the AR(1) model with missing data treated using the OLS and Interpolation with Weighting method on the Bandung City average air temperature data according to the 3-stage procedure of Box Jenkins produces a value of 0.9991 approaching the value of one, this causes the estimated value data is missing a small difference of 0.02126. The comparison of MAPE values for forecasting using the AR(1) model, which has been overcome by the data for the second time gives the same result of 2,459% <10% indicating accurate forecasting.

ACKNOWLEDGEMENT

The authors would like to thank the Rector of Universitas Padjadjaran, who provide financial supports to disseminate research reports of students and lecturers under the Academic Leadership Grant with the grant number: 1959/UN6.3.1/PT.00/2021.

REFERENCES

[1] "Tujuan Pembangunan Berkelanjutan yang Perlu Diketahui oleh Pemerintah Daerah," [Online]. Available: https://www.uclg.org/sites/default/files/tujuan-sdgs.pdf . [Accessed 20 Mei 2021].

[2] D. MenLHK, "Perubahan Iklim," [Online]. Available: http://ditjenppi.menlhk.go.id/kcpi/#:~:text=Perubahan%20Iklim%20adalah%20perubahan%20signifikan,menyebabkan%20efek%20gas%20rumah%20kaca. [Accessed 18 Mei 2021].

[3] C. Sobarna, "Bandung Kota untuk Semua: Harapan dan Tantangan yang Selaras dengan Sustainable Development Goals (SDGs)," *Metahumaniora,* vol. 10, pp. 295-309, 2020.

[4] W. W. S. Wei, Time Series Analysis Univariate and Multivariate Methods Second Edition, USA: Addison-Wesley Publishing Company, 2006.

[5] G. M. J. George E. P. Box, Time Series Analysis Forecasting and Control, California: Holden Day, 1976.

[6] D. S. Fung, *Methods for the Estimation of Missing Values in Time Series,* Australia: Faculty of Communications, Health and Science Edith Cowan University, 2006.

[7] E. T. H. M. Fitriani, "Pemodelan Autoregressive (AR) pada Data Hilang dan Aplikasinya pada Data Kurs Mata Uang Rupiah," *Jurnal Matematika, Statistika, dan Komputasi,* vol. 9, pp. 69-85, 2013.

[8] H. W. Y. W. J. L. H. G. Xi Chen, "Autoregressive Model Based Methods for Online Time Series Prediction with Missing Values: an Experimental Evaluation," *arXiv preprint arXiv,* pp. 2-22, 2019.

[9] O. C. S. S. I. P. Arsali, "Penentuan Koefisien untuk Perhitungan Suhu Udara Rata-rata Harian Data Stasiun Klimatologi Palembang," *Jurnal Meteorologi dan geofisika,* vol. 16, pp. 37-45, 2015.

[10] Suhartono, Analisis Data Statistik dengan R, Surabaya: ITS, 2008.

[11] D. A. Dickey, "Stationarity issues in time series models," *Stationarity issues in time series models. SAS Users Group International, 30,* 2015.

[12] S. C. J. W. Regis Anne, "ARIMA modelling of predicting COVID-19 infections," *medRxiv,* 2020.

[13] A. W. K. M. T. A. S. Fozia Malik, "Box-Cox Transformation Approach for Data Normalization: A Study of New Product Development in Manufacturing Sector Of Pakistan," *IBT Journal of Business Studies (JBS),* vol. 14, no. 1, pp. 110-119, 2018.

[14] C. C. D. T. Eugenio Cesario, “Forecasting Crimes using Autoregressive Models,” in *IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, Italy, 2016.

[15] BMKG, “Data online-Pusat Database,” [Online]. Available: http://dataonline.bmkg.go.id/data_iklim. [Accessed 19 Mei 2021].