BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2165–2180
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2165-2180

PM_{10} AIR QUALITY INDEX MODELING USING ARFIMA-GARCH METHOD: BUNDARAN HI AREA OF DKI JAKARTA PROVINCE

Susilo Hariyanto[1], Salsabila Gustia Wibawa[2*], Solikhin[3]

[1,2,3]*Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro*
Jl. Prof. Soedarto SH, Semarang, 50275, Indonesia

*Corresponding author's e-mail: *salsabilagustiaw@gmail.com*

Article History:

Received: 28th December 2023
Revised: 9th May 2024
Accepted: 5th July 2024
Published: 14th December 2024

Keywords:

ARFIMA-GARCH;
Forecasting;
PM_{10};
Air Quality Index.

ABSTRACT

Air quality is an essential factor in urban life, and its' assessment often relies on the concentration of measurable air pollution parameters. One critical parameter is Particulate Matter (PM), particularly PM_{10}, which comprises solid or liquid particles dispersed in the air from various sources. One of the methods employed for predicting stock index prices is ARFIMA. ARFIMA is used to model long memory data characterized by a slowly decreasing Autocorrelation Function (ACF) plot (hyperbolic) or a difference value in the fractional from. This method is widely used due to its ability to handle nonstationarity issues in time series. However, the time series data often contain heteroskedasticity problems. Data with heteroscedasticity are then further addressed using the GARCH model, because it can model volatility changes occurring over longer periods and capture the persistence of volatility. The ARFIMA-GARCH model can explain long-memory patterns in time series data and address heteroscedasticity issues. The data are sourced from the Jakarta open data web, which is integrated with DLH DKI Jakarta Province. The aim of this research was to forecast the PM_{10} air quality index at the Bundaran HI Area in the Province of DKI Jakarta for the next 14 days, from January 1st to January 14th, 2021, using an ARFIMA model enhanced with GARCH. The analysis reveals that the best model is ARFIMA ([17], d, [1])-GARCH (1,1). Forecasting using this model resulted in a MAPE of 3.47%, indicating that the model is highly capable of forecasting several periods.

How to cite this article:

S. G. Wibawa, S. Hariyanto and Solikhin., "PM10 AIR QUALITY INDEX MODELING USING ARFIMA-GARCH METHOD: BUNDARAN HI AREA OF DKI JAKARTA PROVINCE," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2165-2180, Decmber, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Clean air is a necessity for every living creature to survive. Humans engage in various activities and have a reciprocal relationship with the environment. Therefore, the existence of an index to monitor air pollution is fundamental, and the air pollution standard index is a number that decrees the ambient air quality condition at a specific location. The Air Pollution Standard Index (ISPU) is a unitless number that describes the ambient air quality condition at a specific location based on the impact on human health, aesthetic value, and other living creatures. According to the Minister of Environment Decree Number: KEP 45 / MENLH / 1997 [1], Regarding the Air Pollution Standard Index, the decree is used to provide convenience to the public in obtaining information about ambient air quality at specific locations and times, among other things. In addition, it is also used as a basis for consideration in air pollution control efforts; therefore, it is necessary to prepare an Air Pollution Standard Index.

Forecasting the PM_{10} air quality index is very important in the context of environmental sustainability and human health. In this phenomenon, time series modeling occurs with long memory data characterized by a slowly decreasing (hyperbolic) autocorrelation function (ACF) plot or fractional difference values. Long memory conditions arise in time series data when each observation significantly correlates with other observations, even though the distance between said observations is quite large. Autoregressive Fractional Integrated Moving Average (ARFIMA) modeling can model data with long memory properties [2]. This case can be addressed by integrating additional modeling, specifically using the ARCH (Autoregressive Conditional Heteroscedasticity) or GARCH (Generalized Auto-Regressive Conditional Heteroskedasticity) effect. Conditional heteroscedasticity implies that the residual value of the data depends on the previous residual value. The ARCH model evolved into the GARCH model, incorporating the influence of residuals from both the previous and prior periods [3]. Consequently, combining ARFIMA modeling with the ARCH/GARCH effect facilitates the prediction of long memory characteristics while accommodating data exhibiting signs of heteroscedasticity conditions.

Lintang Furi Prihastari's research demonstrates that the ARIMA-GARCH method effectively forecasts Delta variant COVID-19 cases in Indonesia. By addressing the heteroscedasticity present in the positive cases, the study achieved a high level of forecasting accuracy, with a MAPE of 1.623428% [4]. Moreover, Krzysztof Burnecki and Grzegorz Sikora studied "Identification and validation of a stable ARFIMA process with application to UMTS (Universal Mobile Telekomunikasi System) data" [5] where the ARFIMA model can model UMTS data with long memory conditions in urban areas.

The preceding research states that the ARFIMA model can be used to model long-term time series data and the GARCH model can be used to overcome heteroscedasticity. The ARFIMA-GARCH which combines ARFIMA's ability to model long-term dependencies with GARCH's capability to handle changing volatility, is well-suited for analyzing data like PM_{10}, which exhibits long memory and variability due to external factors such as weather and industrial activity. Despite its potential, this method is rarely applied to environmental data, particularly the PM_{10} air index [6][7]. The author wants to further study the forecasting of PM_{10} air quality index related to the high mobility which is directly proportional to vehicle emissions at the Bundaran HI in DKI Jakarta Province. However, the application of this method to environmental data, especially the PM_{10} air index, is still rarely done. Therefore, this study aims to fill this knowledge gap and provide a better understanding of the patterns and dynamics of the PM_{10} air index at Bundaran HI Area in DKI Jakarta Province.

2. RESEARCH METHODS

The research methodology section provides detailed explanations of the research design, experimental settings, data sources, data collection techniques, and data analysis procedures employed in this study.

2.1 Time Series Analysis

Time series datasets were obtained in order according to time sequence over a certain period. The purpose of implementing time series analysis is to determine the pattern of past data that has been collected based on time sequence for further use in forecasting. Time series forecasting is based on past data behavior

to be projected into the future using mathematical and statistical equations [8]. One of the most important aspects of choosing a suitable forecasting method for time series data is identifying different data patterns.

2.2 Stationarity

A data is called stationarity if the data has not experienced a sharp decline or growth. In other words, the data should generally look horizontal along the time axis. However, it is necessary to differencing data that is still not stationarity to become stationarity. Stationarity data is data that does not contain root units. Stationary data is data that does not contain a unit root where only two-unit root test methods are used, namely:

1. Stationarity in Variance

   It is said to be stationarity in the variance if the rounded value $= 1$, while a Box-Cox transformation can be carried out if it is nonstationarity in the data formulated as follows [9]:

$$T(Z_t) = \begin{cases} \dfrac{Z_t^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln Z_t\,, & \lambda = 0 \end{cases} \tag{1}$$

   Description :

   Z_t : Actual data of period t-th
   λ : Parameters of the transformation

2. Stationarity in Mean

   According to Rosadi [10], non-stationarity in the mean can be observed whether the time series data contains a unit root. One method is the ADF (Augmented Dickey-Fuller) test. The following are common forms of ADF:

$$\Delta(Z_t) = \beta_1 + \beta_2 t + \delta Z_{t-1} + \sum_{i=1}^{n} \alpha_i \Delta Z_{t-1} + a_t \tag{2}$$

   Description :

   $\Delta(Z_t)$: The first difference of the time series Z
   β_1 : The constant value or intercept
   β_2 : The regression coefficient for the trend
   $\sum_{i=1}^{n} \alpha_i \Delta Z_{t-i}$: The sum of lagged differences used to capture serial correlation in the data
   δZ_{t-1} : The regression coefficient for lag Z
   α_t : The residual or random error

   indicated that δ is defined as $\delta = \rho - 1$, where ρ is the coefficient of Z_{t-1}.

2.3 ARFIMA Model

The ARFIMA (Autoregressive Fractionally Integrated Moving Average) model is a development of the ARIMA (Autoregressive Integrated Moving Average) model for modeling Long Memory data [11]. The ARFIMA model is a model that can explain time series data in the form of both short-term and long-term data with differencing (d) fractional values. The general form of the ARFIMA model (p, d, q) is as follows [9][12]:

$$\Phi_p(B)(1 - B)^d Z_t = \theta_q(B) a_t \tag{3}$$

Description:

d : Distinguishing parameters (fractional numbers)
$\Phi_p(B)$: Polynomial autoregressive p-th.
It can be written as: $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$
$\theta_q(B)$: Polynomial moving average q-th.
It can be written as: $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$
B : Backshift operator
a_t : Error or noise at time t, which is usually considered as white noise with zero mean and constant variance.

$(1-B)^d$ as a fractional differentiator operator

For a fractional value of d value, the fractional differentiator operator $(1-B)^d$ is defined as follows:

$$(1-B)^d = 1 + \sum_{k-1}^{\infty} \frac{\Gamma(-d+k)!}{\Gamma(-\mathrm{d})(\mathrm{k})!} B^k \tag{4}$$

where $\Gamma(x)$ is the gamma function.

According to Hosking **[11]** the main characteristics of an ARFIMA(p,d,q) model are as follows:

1. If $|d| \geq 0{,}5$, then the long process is not stationarity.
2. If $0 < d < 0{,}5$, the process correlates stationarity length with the presence of positive dependencies between far-apart observations indicated by positive autocorrelation and slow descent and has a representation of a moving average of infinite order.
3. If $-0{,}5 < d < 0$, then the process correlates stationarity length with having a negative dependence with negative autocorrelation and descending slowly and having an autoregressive representation of infinite order.
4. If $d = 0$, then the process shows the autocorrelation function drops exponentially with the ARMA process.

2.4 Long Memory Scheme

Long memory model identification can also use the Hurst Exponential (H) value. This value can be obtained from the R/S (Rescaled Range Statistic) which is calculated by the following formula **[11]**:

1. Calculating the *mean* $(\bar{Z})$

$$\bar{Z} = \frac{1}{T}\sum_{t=1}^{T} Z_t\,, \quad t = 1{,}2{,}, \dots, T \tag{5}$$

2. Calculating the *adjusted mean*

$$Z_t^{\mathrm{adj}} = Z_t - \bar{Z}, \quad t = 1{,}2{,}, \dots, T \tag{6}$$

3. Calculate cumulative deviation

$$Z_t^* = \sum_{t=1}^{T} Z_t^{\mathrm{adj}} \tag{7}$$

4. Calculate the cumulative deviation range

$$R_t = \max(Z_1^*, Z_2^*, \dots, Z_t^*) - \min(Z_1^*, Z_2^*, \dots, Z_t^*)\,, \quad t = 1{,}2, \dots, n \tag{8}$$

5. Calculating standard deviation

$$\text{the } S_t = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (Z_t - \bar{Z})^2} \tag{9}$$

6. Calculate *rescaled range* (R/S)

$$(R/S)_t = R_t/S_t \tag{10}$$

7. Calculate *log rescaled range statistics* (R/S) values

$$Y_t = \ln[(R/S)_t] \tag{11}$$

8. Calculate time logs from observations

$$X_t = \ln(t) \tag{12}$$

9. Determining the Hurst value

$$H = \frac{\sum_{j=1}^{T}(X_j - \bar{X})\,(Y_j - \bar{Y})}{\sum_{j=1}^{T}(X_j - \bar{X})^2} \tag{13}$$

where $X_j = X_1, X_2, \dots, X_t$ and $Y_j = Y_1, Y_2, \dots, Y_t$

If $H = 0.5$, then the time series is random
If $0 < H < 0.5$, then the short memory process occurs
If $0.5 < H < 1$, then the long memory process occurs

2.5 GARCH Model

The GARCH model was formed to reduce the number of orders that are high enough in the ARCH model because it corresponds to the principle of model selection which is simpler, therefore it will produce variances that are always positive. This parsimony principle also allows GARCH to achieve better predictions with fewer variables and avoid overfitting. General form of GARCH(p, q) [13] :

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{\mathrm{j}=1}^{p} \beta_j \sigma_{t-\mathrm{j}}^2 \tag{14}$$

In the GARCH model, the residual variance of σ_t^2 is not only influenced by the residual of the past period (ε_{t-i}^2), but also the residual variant of the past period (σ_{t-j}^2) where:

$$p \geq 0; q \geq 0$$
$$\omega > 0, \alpha_i \geq 0, i = 1,2, \dots, q$$
$$\beta_j \geq 0, j = 1,2, \dots, p$$

Description:

σ_t^2 : Variance of the residual at time t
ω : Constant components
α_i : Parameter coefficient of ARCH order p
β_j : Parameter coefficients of GARCH order q
ε_{t-i}^2 : Residual squares in the period $t - i$

2.6 Best Model Selection

AIC (Akaike's Information Criterion) posits that a model with a lower AIC value is better as it indicates a balance between model goodness and model complexity which suggests the model is more capable of generalizing to new data. In model selection, the computed AIC value is used [9].

$$AIC = n \ln \hat{\sigma}_a^2 + \frac{2k}{n}. \tag{15}$$

Description:

$\hat{\sigma}_a^2$: Maximum likelihood estimation of σ_a^2
k : The number of parameters in the model
n : The number of observations

2.7 Selection of the Best Forecasting Accurancy

Selection of the best forecasting accuracy through measuring the level of accuracy using Mean Absolute Percentage Error (MAPE), where the smaller the MAPE value, the smaller the forecasting error, and the better the forecasting results will be closer to the actual value. MAPE can be calculated using the following formula [14] :

$$\text{MAPE} = \left[\frac{1}{n}\sum_{t=1}^{n} \frac{|Z_t - \hat{Z}_t|}{Z_t}\right] \times 100\%. \tag{16}$$

Description:

n : Number of forecasting periods

Z_t : Actual data values in period t
$\hat{Z}_t$: Forecast value in period t

3. RESULTS AND DISCUSSION

3.1 Missing Values

The integrity of data analysis outcomes can be compromised by missing values, potentially leading to reduced accuracy stemming from non-sampling errors. To effectively address these concerns, it's crucial to categorize missing values into distinct classifications. Missing Completely at Random (MCAR) values share no relationship with other variables, while Missing at Random (MAR) values exhibit a connection to other variables, enabling estimation through examination of these relationships. In contrast, Missing Not at Random (MNAR) values demonstrate a relationship with other missing values, rendering estimation through existing variables infeasible. To mitigate the impact of missing data, linear interpolation represents a viable approach for filling in these gaps [15][16][17].

3.2 Linear Interpolation

Linear interpolation is a polynomial of the first degree and through a straight line at every two successive input points where interpolating two points with a straight line using two pairs of points to obtain a range of values $y = f_1(x)$ if know x is between x_o and x_1 [10][18] Application example:

Table 1. Data That Contains Missing Value

Date	PM_{10}
27/01/2019	24
28/01/2019	-
29/01/2019	-
30/01/2019	52

Data Source: *https://data.jakarta.go.id/*

Table 2. Data That Has Been Done Linear Interpolation

Date	PM_{10}
27/01/2019	24
28/01/2019	9.333
29/01/2019	33.333
30/01/2019	52

Data Source: *Rapidminer Software Program, 2023*

Step 1: find the value of m on 01/27/2019

By assuming the date 27/01/2019 as $x_0 = 1$ and the date 30/01/2019 as $x_1 = 4$

$$m = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{52 - 24}{4 - 1} = \frac{28}{3} = 9.333$$

Step 2: search for missing values on 28/01/2019

Suppose the date 28/01/2019 as $x = 2$

$$f_1(x) - f(x_0) = m(x - x_0)$$
$$f_1(x) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$
$$f_1(x) - f(x_0) = (9{,}333)(2 - 1)$$
$$f_1(x) = (9{,}333) + f(x_0) = (9{,}333) + 24 = 33.333$$

Thus, for blank value dated 29/01/2019

$$f_2(x) = (9{,}333) + f_1(x)$$
$$f_2(x) = (9{,}333) + 33.333 = 42.666 \approx 42.7$$

Therefore, the estimated approach figures for PM_{10} for 28/01/2019 = 33.3 and 29/10/2019 = 42.7 as in **Table 2.**

3.3 Data Description

The study analyzes 731 days of daily PM_{10} data from Jakarta (2019-2020), sourced from jakartaopendata and DLH DKI Jakarta Province, and visualized using R studio. The data, visualized in **Figure 1**, shows PM_{10} movement before processing and analysis.

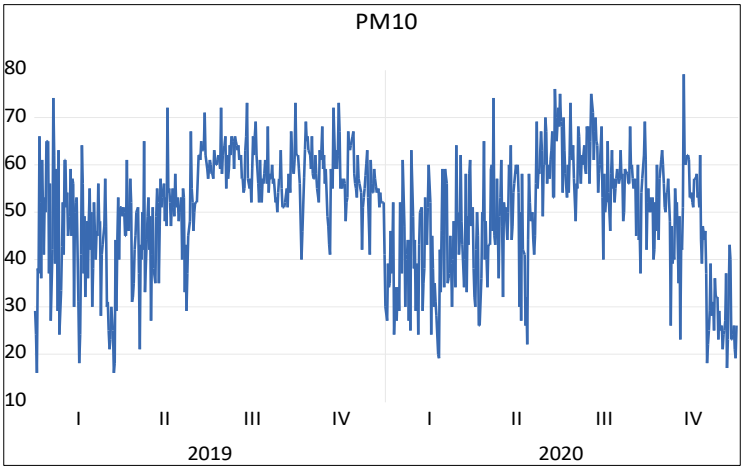
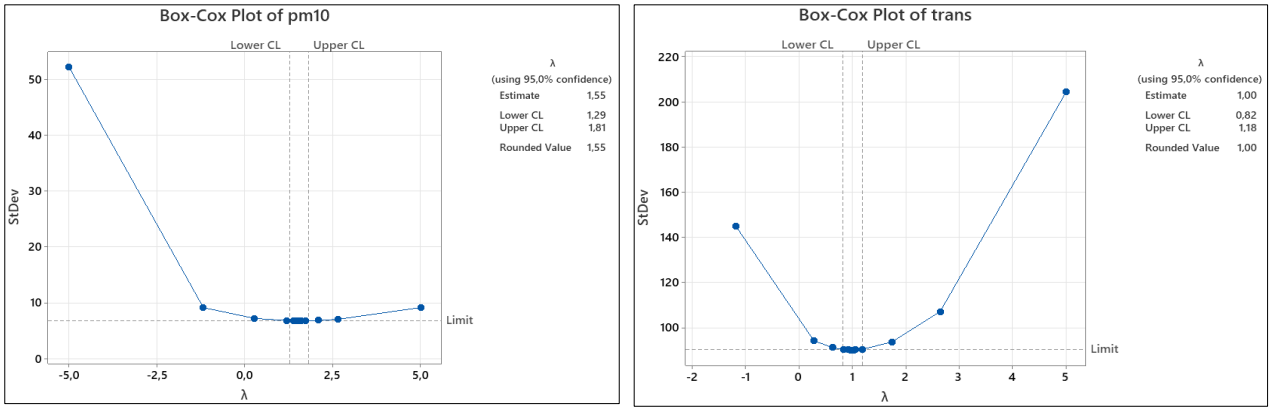


Figure 1. Graph of PM_{10} Air Quality Index for the Period of January 1, 2019 - December 31, 2020

3.4 Stationarity

3.4.1 Stationarity in Variance

Testing stationarity data in a variance using the Box-Cox test, if the data has been stationarity in the rounded variance value = 1. If the rounded value is not equal ≠ 1 , the data needs to be transformed.



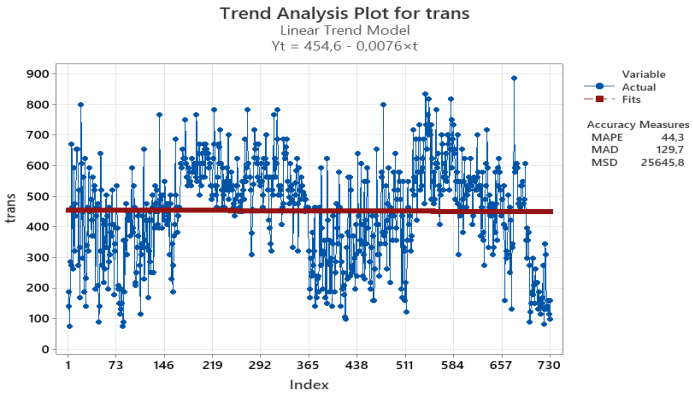
(a) (b)

Figure 2. Stationarity In Variance (a) Before Transformation; (b) After Transformation

Based on **Figure 2** (a) shows PM_{10} data is not stationarity in variance ($\lambda = 1.55$), and **Figure 2** (b) confirms the data has been stationarity in the variance because rounded value = 1.

3.4.2 Stationarity in Mean

Two methods can be used to test stationarity data in the mean: visual and formal tests. Visual tests can be seen by observing the data plotted against time, and if the plot is not far from the mean value and does not show symptoms of a trend, then stationarity data is in the mean visually.



(a)

Null Hypothesis: PM10 has a unit root
Exogenous: Constant
Lag Length: 4 (Automatic - based on SIC, maxlag=19)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-5.259270	0.0000
Test critical values:	1% level	-3.439130	
	5% level	-2.865305	
	10% level	-2.568831	

*MacKinnon (1996) one-sided p-values.

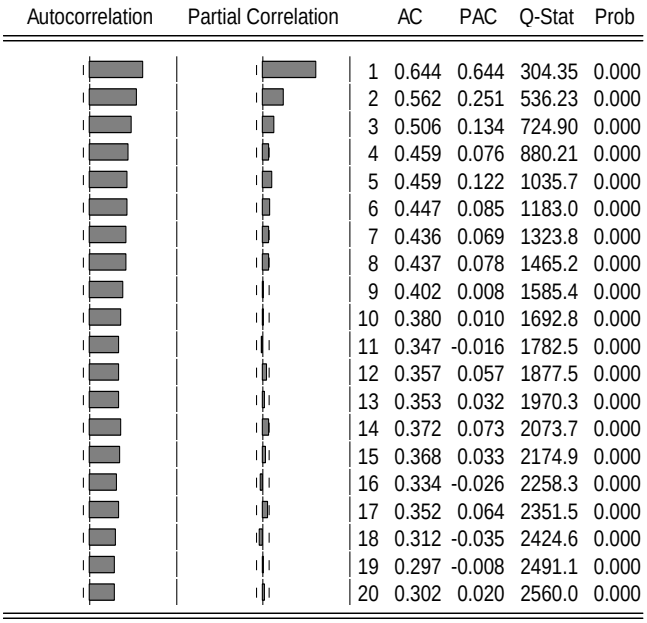
(b)

Figure 3. Trend Time Series PM_{10} Air Quality Index Data (a) Plot, (b) ADF Test

The transformed PM_{10} data, as shown in **Figure 3** (a), indicates stationarity in the mean, characterized by consistent fluctuations around the mean without a unit root. The ADF test (**Figure 3** (b)) yields a p-value of $0.0000 < \alpha = 0.05$, confirming that the daily PM_{10} data is stationarity without requiring differencing.

3.4.3 Correlogram Stationarity Test

Data that had previously been known to be stationarity was then checked by the correlogram. Model identification is carried out by making an ACF plot to find out the q order of the MA(q) model and a PACF plot to find out the p order of the AR(p) model to find out the lag visually.



	AC	PAC	Q-Stat	Prob
1	0.644	0.644	304.35	0.000
2	0.562	0.251	536.23	0.000
3	0.506	0.134	724.90	0.000
4	0.459	0.076	880.21	0.000
5	0.459	0.122	1035.7	0.000
6	0.447	0.085	1183.0	0.000
7	0.436	0.069	1323.8	0.000
8	0.437	0.078	1465.2	0.000
9	0.402	0.008	1585.4	0.000
10	0.380	0.010	1692.8	0.000
11	0.347	-0.016	1782.5	0.000
12	0.357	0.057	1877.5	0.000
13	0.353	0.032	1970.3	0.000
14	0.372	0.073	2073.7	0.000
15	0.368	0.033	2174.9	0.000
16	0.334	-0.026	2258.3	0.000
17	0.352	0.064	2351.5	0.000
18	0.312	-0.035	2424.6	0.000
19	0.297	-0.008	2491.1	0.000
20	0.302	0.020	2560.0	0.000

Figure 4. ACF and PACF Correlogram Plot

Based on **Figure 4** on the ACF plot pattern shows that the ACF plot lag drops hyperbolically which means it contains a Long Memory effect. For this reason, it is necessary to identify the effect of Long Memory. If the data contains a Long Memory effect, the next step is differentiating using ARFIMA while if there is no Long Memory effect continued with the ARIMA method.

3.5 Identify Long Memory Patterns

Testing of the Long Memory effect is formally carried out by calculating the Hurst value obtained by logarithmic R/S statistics and estimating it by the Ordinary Least Square (OLS) method. Accordingly, the following results are obtained:

1. Calculating the *mean* $(\bar{Z})$ transformed PM_{10} air quality index data

$$\bar{Z} = \frac{1}{T}\sum_{t=1}^{T} Z_t = \frac{330243}{731} = 451.768$$

2. Calculating the *adjusted mean* of each PM_{10} data

$$Z_1^{\text{adj}} = Z_1 - \bar{Z} = 186{,}602 - 451{,}768 = -265.166$$
$$Z_2^{\text{adj}} = Z_2 - \bar{Z} = 139{,}088 - 451{,}768 = -312.680$$
$$\vdots$$
$$Z_{731}^{\text{adj}} = Z_{731} - \bar{Z} = 157{,}497 - 451{,}768 = -294.271$$

3. Calculate the cumulative deviation or standard deviation

$$S_1 = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Z_1 - \bar{Z})^2} = 265.166$$
$$S_2 = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Z_2 - \bar{Z})^2} = 221.098$$
$$\vdots$$
$$S_{731} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Z_t - \bar{Z})^2} = 10.884$$

4. Calculating cumulative deviation range PM_{10} data

$$Z_1^* = \sum_{t=1}^{T} Z_t^{\text{adj}} = Z_1^{\text{adj}} = -256.166$$
$$Z_2^* = \sum_{t=1}^{T} Z_t^{\text{adj}} = Z_1^{\text{adj}} + Z_2^{\text{adj}} = -577.846$$
$$\vdots$$
$$Z_{731}^* = \sum_{t=1}^{T} Z_t^{\text{adj}} = Z_1^{\text{adj}} + Z_2^{\text{adj}} + \cdots + Z_{731}^{\text{adj}} = 0$$

5. Calculating standard deviation

$$R_1 = \max(Z_1^*) - \min(Z_1^*) = -256{,}166 - (-256{,}166) = 0$$
$$R_2 = \max(Z_1^*, Z_2^*) - \min(Z_1^*, Z_2^*) = -256{,}166 - (-577{,}846) = 312.680$$
$$\vdots$$
$$R_{731} = \max(Z_1^*, Z_2^*, \dots, Z_{731}^*) - \min(Z_1^*, Z_2^*, \dots, Z_{731}^*) = 20855{,}689$$

6. Calculate rescaled range (R/S)

$$(R/S)_1 = R_1/S_1 = 0$$
$$(R/S)_2 = R_2/S_2 = 1.414$$
$$\vdots$$
$$(R/S)_{731} = R_{731}/S_{731} = 1916{,}177$$

7. Calculate $log\ rescaled\ range\ statistics$ (R/S) values

$$Y_1 = \ln[(R/S)_1] = 1$$
$$Y_2 = \ln[(R/S)_2] = 0.347$$
$$\vdots$$
$$Y_{731} = \ln[(R/S)_3] = 7.558$$

8. Calculate time logs from observations

$$X_1 = \ln(1) = 0$$
$$X_2 = \ln(2) = 0.693$$
$$\vdots$$
$$X_{731} = \ln(731) = 6.594$$

9. Determine the value of Hurst.

$$H = \frac{\sum_{j=1}^{T}(X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^{T}(X_j - \bar{X})^2} = 0.8844464$$

Based on the Hurst Exponent (H) value that has been done which is 0.8844464 because the Hurst Exponent value is located in the interval $0.5 < H < 1$, it can be concluded that there is long memory in the data.

3.6 ARFIMA Modeling

3.6.1 Estimation of The Distinguishing Parameters (d)

Long Memory processes are marked with a differentiating value (d) in fractional form. The method used to determine the value of the distinguishing parameter (d) by the Geweke Porter-Hudak (GPH) estimator [11][19]. The estimated value of d obtained from R Studio software is $d = 0.4905795$.

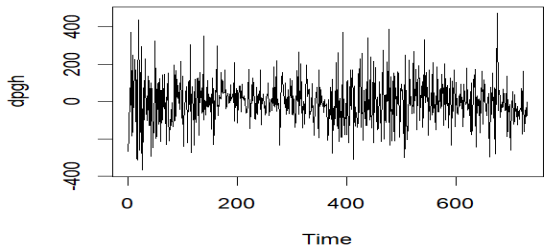


Figure 5. Time Series Plot Data Diff(d)

Hipotesis
$H_0: \delta = 1$ (Data is not stationarity or there is a unit root)
$H_1: \delta < 1$ (Stationarity data or no root unit)

Figure 5 is PM_{10} air quality index data at Bundaran HI Area in DKI Jakarta Province after differencing with GPH estimates shows that the differencing data does not form an upward or downward trend and confirms the data is no unit root or has been stationarity in the mean ($p - value = 0.01;\ \alpha = 5\%$).

3.6.2 ARFIMA Model Identification

ARFIMA model identification can be determined based on ACF and PACF plots. ACF plots are used to identify MA models and PACF is used to use AR models. Based on the results of R studio, the following results are obtained:

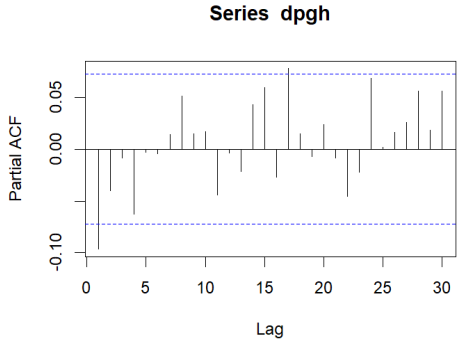


Figure 6. Plot PACF Data Diff(d)

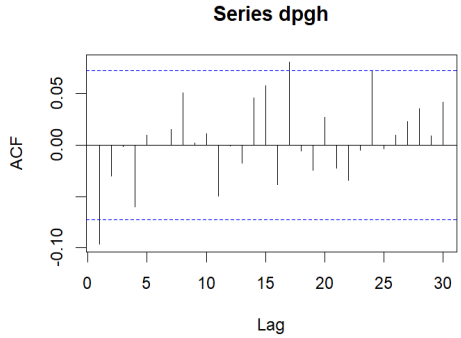


Figure 7. Plot ACF Data Diff(d)

Figure 6 and Figure 7 show that the PACF and ACF plots have lag values in the PACF plot: AR(p) = 1 and 17 and the ACF plot: MA(q) = 1, 17, and 24 where the lag values will form 11 models which will then be estimated the parameters and significance of the ARFIMA model.

3.6.3 Parameter Estimation and Significance of ARFIMA Model

The parameter significance test is useful to ensure that the resulting ARFIMA model has a significant contribution in explaining the variation of time series data. The hypothesis test in estimating ARFIMA model parameters is as follows:

H_0: parameter = 0 (parameter is not significant to the model)
H_1: parameter ≠ 0 (parameter significant to model)
The significance level is alpha is $\alpha = 0{,}05$

Test criteria : Reject H_0 if $|t_{\text{count}}| \geq \frac{t_\alpha}{2}; n-1$ or if $p - value < \alpha$; where n = number of observations

$$\frac{t_\alpha}{2} = \frac{\alpha}{2} = 0.025; n - 1 = 731 - 1 = 730$$

$$t(0{,}025; 730) = 1.963219$$

The estimation of the parameters and significance of the ARFIMA model there are 8 significant models from 11 ARFIMA models, as presented in **Table 3.**

Table 3. Parameter Estimation and Significance of ARFIMA Model

Model	Parameter	Coefficient	T-statistic	T-table	p-value	Decision	Description
ARFIMA ([1], d, 0)	ϕ_1	-0.0968	-2.6230	1.9632	0.0087	Declined	Significant
ARFIMA ([17], d, 0)	ϕ_{17}	0.0855	2.2623	1.9632	0.0237	Declined	Significant
ARFIMA (0, d, [1])	θ_1	-0.1047	-2.7318	1.9632	0.0063	Declined	Significant
ARFIMA (0, d, [17])	θ_{17}	0.07670	2.1548	1.9632	0.0312	Declined	Significant
ARFIMA (0, d, [24])	θ_{24}	0.0737	1.9929	1.9632	0.0463	Declined	Significant
ARFIMA ([1], d, [17])	ϕ_1	-0.0950	-2.5725	1.9632	0.0101	Declined	Significant
	θ_{17}	0.0749	2.0916	1.9632	0.0365	Declined	Significant
ARFIMA ([1], d, [24])	ϕ_1	-0.0975	-2.6409	1.9632	0.0083	Declined	Significant
	θ_{24}	0.0750	2.0166	1.9632	0.0437	Declined	Significant
ARFIMA ([17], d, [1])	ϕ_{17}	0.0831	3.1982	1.9632	0.0279	Declined	Significant
	θ_1	-0.1030	-2.6771	1.9632	0.0074	Declined	Significant

Data Source: *R Studio software program, 2023*

3.7 Diagnostic Tests

3.7.1 Residual Independence Test

The assumption of residual independence to measure the linear relationship between lag values in time series data. The residual independence test is also called the autorelation test or white noise test. The residual independence assumption is satisfied if the residual is white noise or there is no correlation between the residual lag **[3][9]**. The hypothesis:

H_0 : There is no correlation between residual lag
H_1 : There is a correlation between residual lag

Test statistic: $LB = n(n+2)\sum_{k=1}^{m}\left(\frac{\hat{\rho}_k^2}{n-k}\right)$; Test criteria: H_0 reject if $LB > \chi^2_{\alpha;2}$ or $p - value < \alpha$

The residual independence test confirms that all ARFIMA models have independent residuals, as supported by significant results at the $\alpha = 5\%$ level.

Table 4. ARFIMA Residual Deficiency Assumption Test

Model	Lag	Q	$p - value$	Decision
ARFIMA ([1]. d. 0)	10	6.5230	0.7696	Accepted
	20	19.0002	0.5218	Accepted
	30	27.9790	0.5716	Accepted
ARFIMA ([17]. d. 0)	10	12.0560	0.2813	Accepted
	20	20.2251	0.4439	Accepted
	30	28.2558	0.5569	Accepted
ARFIMA (0. d. [1])	10	6.1243	0.8047	Accepted
	20	18.6056	0.5476	Accepted
	30	27.7240	0.5851	Accepted
ARFIMA (0. d. [17])	10	12.0947	0.2788	Accepted
	20	20.2081	0.4450	Accepted
	30	28.2061	0.5595	Accepted
ARFIMA (0. d. [24])	10	12.9865	0.2244	Accepted
	20	26.3037	0.1560	Accepted
	30	30.8367	0.4235	Accepted
ARFIMA ([1]. d. [17])	10	6.4643	0.7749	Accepted
	20	15.0503	0.8204	Accepted
	30	23.3848	0.7991	Accepted

Model	Lag	Q	$p-value$	Decision
ARFIMA ([1]. d. [24])	10	6.8586	0.7387	Accepted
	20	19.2772	0.5039	Accepted
	30	24.7057	0.7392	Accepted
ARFIMA ([17]. d. [1])	10	6.0337	0.8124	Accepted
	20	14.4378	0.8076	Accepted
	30	23.1572	0.8087	Accepted

Data Source: *R Studio software program, 2023*

3.7.2 Residual Normality Test

The normality test is carried out to see the normality of the residual data as a result of the transformation. Formally, a normal distribution assumption test was carried out using the Kolmogorov-Smirnov test. The stages of the hypothesis are as follows:

$H_0: F(x) = F_0(x)$ (Residual normally distributed data)
$H_1: F(x) \neq F_0(x)$ (Residual data is not normally distributed)

The level of significance is $\alpha = 0.05$; Test statistic : $D = Sup|S(x) - F_0(x)|$; Test criteria: H_0 reject if $D > K_{(1-\alpha);n}$ or $p - value < \alpha$

Table 5. Kolmogorov-Smirnov ARFIMA Test

Model	$p-value$	Decision
ARFIMA ([1], d, 0)	0.0006	Declined
ARFIMA ([17], d, 0)	0.0007	Declined
ARFIMA (0, d, [1])	0.0006	Declined
ARFIMA (0, d, [17])	0.0006	Declined
ARFIMA (0, d, [24])	0.0002	Declined
ARFIMA ([1], d, [17])	0.0017	Declined
ARFIMA ([1], d, [24])	0.0006	Declined
ARFIMA ([17], d, [1])	0.0022	Declined

Data Source: *R Studio software program, 2023*

Based on **Table 5** the ARFIMA models formed residual values are not normally distributed because all the $p - value < \alpha = 0.05$. One of the assumptions that must be met in the model is that residuals of data are normally distributed. However, according to Rosadi **[10]**, it is said that the assumption of normality is less important than the assumption of independence therefore, the assumption of normality can be ignored.

3.7.3 Residual Heteroscedasticity Test

The residual heteroscedasticity assumption test is used to determine whether the ARFIMA model obtained has a constant error variant or not to get the ARCH / GARCH effect on residuals. Non-constant errors mean that the model has heteroskedasticity problems. Heteroskedasticity test on ARFIMA model using ARCH-LM method.

Test statistic : $LM = \frac{SSR_0}{SSR_1/\left(\frac{n-2u-1}{u}\right)}$; with $SSR_0 = \sum_{t=u+1}^{n}(a_t^2 - \bar{a}_t^2)^2$, $\bar{a}_t^2 = \frac{\sum_{t=1}^{n} a_t^2}{n}$, $SSR_1 = \sum_{t=u+1}^{n} \bar{a}_t^2 \, e_t^2$.

Indicated e_t^2 the smallest residual quarter of the equation and u is specified by a positive integer.

Based on **Table 6** of the significant level of $\alpha = 5\%$, it can be concluded that the ARFIMA models above have an ARCH/GARCH effect on residual data.

Table 6. ARFIMA Residual Heteroscedasticity Test

Model	LM	$p-value$	Decision
ARFIMA ([1], d, 0)	34.555	0.0005	Declined
ARFIMA ([17], d, 0)	33.964	0.0007	Declined
ARFIMA (0, d, [1])	34.179	0.0006	Declined
ARFIMA (0, d, [17])	34.41	0.0006	Declined
ARFIMA (0, d, [24])	37.892	0.0002	Declined
ARFIMA ([1], d, [17])	31.46	0.0017	Declined
ARFIMA ([1], d, [24])	34.535	0.0006	Declined
ARFIMA ([17], d, [1])	30.691	0.0022	Declined

Data Source: *R Studio software program, 2023*

3.8 Best Model Selection

The best model selection among eight model is ARFIMA ([17], d, [1]) with an AIC values of 9019.857 where the ARFIMA model equation ([17], d, [1]) is as follows:

$$\Phi_p(B)(1-B)^d Z_t = \theta_q(B)a_t$$
$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_{17}B^{17})(1-B)^d Z_t = (1-\theta_1 B)a_t$$
$$(1-\phi_1 B)(1-B)^d Z_t = a_t-\theta_1 a_{t-1}$$
$$(1-0.083143B)(1-B)^{0.4905795} Z_t = a_t+0.10303a_{t-1}$$

3.9 GARCH Modeling

3.9.1 Parameter Estimation and Significance Test

Hypothesis testing of parameter estimates for the ARFIMA-GARCH models showed significant results at the α = 0.05 level for four specific model configurations. Notably, all parameters within these four models demonstrated statistical significance, underlining their relevance to the overall model structure.

Table 7. Results of Parameter Estimation and ARFIMA-GARCH Significant Test

Model	Para-meter	Parameter Estimation	T -Statistic	T-table	$P-value$	Decision	Description
ARFIMA ([17], 0, d, [1])-GARCH(0,1)	β_1	0.9848	1.73×10^{-2}	1.9632	0	Declined	Significant
ARFIMA ([17], 0, d, [1])-GARCH(1,0)	α_1	0.5021	5.9×10^{-3}	1.9632	0	Declined	Significant
ARFIMA ([17], 0, d, [1])-GARCH(2,0)	α_2	0.6921	3119.1	1.9632	0	Declined	Significant
ARFIMA ([17], 0, d, [1])-GARCH(1,1)	α_1	0.0352	3.299883	1.9632	0.0010	Declined	Significant
	β_1	0.9508	71.17763	1.9632	0	Declined	Significant

Data Source: *R Studio software program, 2023*

3.9.2 Best Model Selection

The smallest value of the four ARFIMA-GARCH models is found in the ARFIMA ([17], d, [1])-GARCH(1,1) model with an AIC value of 12.273 therefore the test will continue using the models.

Table 8. Best Model Selection

Model	AIC
ARFIMA ([17], 0, d, [1])-GARCH(0,1)	12.315
ARFIMA ([17], 0, d, [1])-GARCH(1,0)	74.281
ARFIMA ([17], 0, d, [1])-GARCH(2,0)	14.258
ARFIMA ([17], 0, d, [1])-GARCH(1,1)	12.273

Data Source: *R Studio software program, 2023*

3.9.3 Sign and Size Bias Test GARCH Model

The ARFIMA model ([17], d, [1])-GARCH(1,1) was then tested for sign and size bias to determine whether the effect model was asymmetric. The sign and size bias test consists of three tests, namely the sign of bias test, the positive bias size test and the negative bias size test expressed in the regression equation which is carried out simultaneously called the joint effect [20]. The stages of the hypothesis are as follows:

$H_o: b_j = 0$ (Residuals are symmetric)
$H_1: b_j \neq 0$ (Residuals are not symmetrical

The level of significance is $\alpha = 0.05$. Test statistic: $t = \frac{\overline{b_j}}{se(\overline{b_j})}$;); $j = 1,2,3$.

The outcomes for the ARFIMA([17], d, [1])-GARCH(1,1) model, as generated using R Studio software [19], are comprehensively presented in Table 9 where it has a calculated t value smaller than the T-table and a $p-value$ more than 0.05. Then it can be decided that the residuals of the model are symmetrical.

Table 9. Sign and Size Bias Test Results of ARIMA-GARCH Model

Test	T-count	T-table	$p-value$	Decision	Description
Sign Bias	0.9720	1.9632	0.3314	H_o Accepted	Symmetrical Residuals
Negative Sign Bias	0.5806	1.9632	0.5617	H_o Accepted	Symmetrical Residuals
Positive Sign Bias	0.7580	1.9632	0.7821	H_o Accepted	Symmetrical Residuals
Joint Effect	1.0792	1.9632	0.0782	H_o Accepted	Symmetrical Residuals

Data Source: *R Studio software program, 2023*

3.9.4 Non Heteroskedasticity Test

It can be concluded that the ARFIMA model ([17], d, [1])-GARCH(1,1) has fulfilled the assumption of independence and there is no heteroscedasticity effect because it has a $p-value = 1$ greater than 0.05. Therefore, obtained the final model of ARFIMA ([17], d, [1])-GARCH(1,1) with the following equation:

$$\Phi_p(B)(1-B)^d Z_t = \theta_q(B)a_t$$
$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_{17}B^{17})(1-B)^d Z_t = (1-\theta_1 B)a_t$$
$$(1-\phi_1 B)(1-B)^d Z_t = a_t-\theta_1 a_{t-1}$$
$$(1-0.045759B)(1-B)^{0.4905795} Z_t = a_t+0.245419a_{t-1},$$

with

$$a_t \sim N(0,\sigma_t^2),$$

and its' variant models

$$\sigma_t^2 = 161.079924+0.035192\varepsilon_{t-1}^2+0.950813\sigma_{t-1}^2.$$

3.10 Prediction Results

The prediction result of the best model ARFIMA ([17], d, [1])-GARCH(1,1), for the next 14 days (January 1st - January 14th, 2021) because air quality index data can change at any time.

Table 10. ARFIMA-GARCH Forecasting Results and Out Sample Data

Period	Date	Predict Data Before Retransformation	Predict Data After Transformation	*Out Sample* Data
1	01/01/2021	153.7	25.7	38
2	02/01/2021	151.7	25.5	27
3	03/01/2021	152.6	25.6	44
4	04/01/2021	145.3	24.8	30
5	05/01/2021	145.7	24.9	38
6	06/01/2021	160.1	26.4	41
7	07/01/2021	165.8	27.0	35
8	08/01/2021	156.2	26.0	37
9	09/01/2021	155	25.9	47
10	10/01/2021	160.8	26.5	23
11	11/01/2021	154.6	25.8	38
12	12/01/2021	150	25.3	29
13	13/01/2021	145.3	24.8	34
14	14/01/2021	147.2	25.0	36

Data Source: *R Studio software program, 2023*

To assess the forecasting accuracy in **Table 10**, using the MAPE function measures the percentage error between the out sample data and the forecasted data in the ARFIMA ([17], d, [1])- GARCH (1,1) model which involves comparing the absolute difference between the actual value and the predicted value, then averaging these values and expressing the result as a percentage of the actual value. To help determine the model's performance in predicting future data points, providing insight into its reliability and accuracy.

$$MAPE = \left[\frac{1}{n}\sum_{t=1}^{n}\frac{|Z_t - \hat{Z}_t|}{Z_t}\right] \times 100\%$$

$$= \frac{1}{14}\left(\left|\frac{38-25.7}{38}\right| + \left|\frac{27-25.5}{27}\right| + \cdots + \left|\frac{36-25}{36}\right|\right) \times 100\% = 0.2724\%$$

The MAPE value obtained shows that the ARFIMA ([17], d, [1])-GARCH(1,1) model has quite good forecasting capabilities.

4. CONCLUSIONS

The analysis of PM10 air quality data at the Bundara HI, DKI Jakarta Province using the ARFIMA([17], d, [1])-GARCH(1,1) model has shown quite good results. This is particularly relevant since the training data covers the period from early 2019 to the end of 2020, including both the pre-pandemic phase and the pandemic period, when there was a significant decrease in vehicle volume due to COVID-19 restrictions in Indonesia. However, this model could be improved by including external variables such as the impact of the COVID-19 pandemic, changes in mobility policies, the level of public compliance with restrictions, and changes in transportation patterns that directly affect vehicle emissions and air quality. Additionally, considering alternative models like ARMAX-GARCH, VARMA-GARCH, SARIMA, or LSTM might be beneficial.

REFERENCES

[1] M. N. Lingkungan Hidup, *Keputusan Menteri Negara Lingkungan Hidup No. 45 Tahun 1997 Tentang : Indeks Standar Pencemar Udara*. 1997.

[2] P. Kartikasari, H. Yasin, and D. A. I. Maruddani, "Autoregressive Fractional Integrated Moving Average (ARFIMA) Model To Predict Covid-19 Pandemic Cases In Indonesia," *MEDIA STATISTIKA*, vol. 14, no. 1, pp. 44–55, Jun. 2021, doi: 10.14710/medstat.14.1.44-55.

[3] R. S. Tsay, *Analysis of Financial Time Series Second Edition*. New Jersey: John Wiley & Sons, Inc, 2005.

[4] L. F. Prihastari, "Mengonstruksi Model Peramlaan Kasus Positif Covid-19 Varian Delta Menggunakan Model ARIMA-GARCH," 2022.

[5] K. Burnecki and G. Sikora, "Identification and validation of stable ARFIMA processes with application to UMTS data," *Chaos Solitons Fractals*, vol. 102, pp. 456–466, Sep. 2017, doi: 10.1016/j.chaos.2017.03.059.

[6] I. M. Ghani and H. A. Rahim, "Modeling and Forecasting of Volatility using ARMA-GARCH: Case Study on Malaysia Natural Rubber Prices," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Aug. 2019. doi: 10.1088/1757-899X/548/1/012023.

[7] T. Rizqiyah and I. Rosyida, "Analisis Cluster Tingkat Kualitas Udara Ambien Jalan Raya di Jawa Tengah Tahun," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 4, pp. 560–564, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[8] S. Makridakis, Wheelwright, and McGee, *Metode dan Aplikasi Peramalan Edisi Kedua*. Jakarta: Erlangga, 1999.

[9] W. W. S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed. New Jersey: Pearson Prentice Hall, 2006.

[10] D. Rosadi, *Analisis ekonometrika & runtun waktu terapan dengan R : aplikasi untuk bidang ekonomi, bisnis, dan keuangan*. Yogyakarta: Andi, 2011.

[11] J. R. M. Hosking, "Fractional Differencing," *Biometrika*, vol. 68, no. 1, p. 165, Apr. 1981, doi: 10.2307/2335817.

[12] N. Rismawati and S. Sugiman, "Long Memory Volatility Model dengan ARFIMA-HYGARCH Untuk Meramalkan Return Indeks Harga Saham Gabungan (IHSG)," *Unnes Journal of Mathematics*, vol. 11, no. 1, pp. 80–91, 2022, doi: 10.15294/ujm.v9i1.36464.

[13] E. M. Y. Wu and S. L. Kuo, "Air quality time series based GARCH model analyses of air quality information for a total quantity control district," *Aerosol Air Qual Res*, vol. 12, no. 3, pp. 327–339, Jun. 2012, doi: 10.4209/aaqr.2012.03.0051.

[14] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed. United States of America: Spinger, 2002.

[15] P. Schmitt, J. Mandel, and M. Guedj, "A Comparison of Six Methods for Missing Data Imputation," *J Biomet Biostat*, vol. 6, no. 1, pp. 1–6, 2015, doi: 10.4172/2155-6180.1000224.

[16] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *Springer Plus*, pp. 1–17, 2013, Accessed: Jun. 16, 2023. [Online]. Available: http://www.springerplus.com/content/2/1/222

[17] E. Hartini, "IMPLEMENTATION OF MISSING VALUES HANDLING METHOD FOR EVALUATING THE SYSTEM/COMPONENT MAINTENANCE HISTORICAL DATA," *J. Tek. Reaktor. Nukl*, vol. 19, no. 1, pp. 11–18, 2017.

[18] Solikhin and S. Khabibah, *Buku Ajar Metode Numerik*. Semarang: UPT UNDIP Press Semarang, 2014.

[19] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications With R Examples*, 2nd ed. New York: Springer, 2006.

[20] Y. Hong and Y.-J. Lee, "A General Approach to Testing Volatility Models in Time Series," *JMSE*, vol. 2017, no. 1, pp. 1–33, 2017, doi: 10.3724/SP.J.1383.201001.