EQA - International Journal of Environmental Quality
ISSN 2281-4485 - Vol. 65 (2025): 86-96
Journal homepage: https://eqa.unibo.it/

published by

Forecasting air quality index data with autoregressive integrated moving average models

Arie Vatresia[1], Ridha Nafila[1], Winalia Agwil[2], Ferzha Putra Utama[3], Maryam Shehab[4]

[1] Department of Informatics, Faculty of Engineering, University of Bengkulu, Indonesia
[2] Department of Information System, Faculty of Engineering, University of Bengkulu, Indonesia
[3] Department of Statistic, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Indonesia
[4] Environmental Protection Authority (EPA), Shuwaikh, Kuwait City, Kuwait

* Corresponding author E-mail: avatresia@unib.ac.id

Article info
Received 2/9/2024; received in revised form 25/9/2024; accepted 7/10/2024
DOI: 10.6092/issn.2281-4485/20263

Abstract
Air pollution arises from several sources, encompassing industrial, transportation, and home activities, and carries significant implica- tions for environmental health. High population mobility in a place, such as Jakarta, might exacerbate air pollution. In 2021, Jakarta, des- ignated as the Special Capital Region, had the highest population den- sity in Indonesia, with 15,978 individuals per square kilometer (km^2). IQAir reports that Jakarta frequently places among the cities with the most unfavorable air quality globally. In 2021, Jakarta was identified as the most polluted city in Indonesia, while Indonesia was placed 17th out of 118 countries for having the poorest air quality. Hence, the Jakarta Environmental Agency has formulated an Air Pollution Control Strategy till 2030 to diminish the proportion of lethal pollu- tion levels. Given the significance of air pollution's detrimental effects on health, it is imperative to consistently regulate and oversee air pol- lution, including forecasting. This study utilizes the forecasting of the Air Quality Index (AQI) in Jakarta at air quality monitoring stations DKI1, DKI2, DKI3, DKI4, and DKI5. The Air Quality Index (AQI) data for Jakarta were obtained from the Jakarta Open Data portal spanning the years 2010 to 2021. The ARIMA model was utilized to process this data. The generated models were assessed for error levels using the parameters Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Ab- solute Percentage Error (MAPE). This study produced a total of 25 ARIMA models to forecast the levels of air quality index (AQI) contaminants. The levels of PM_{10}, SO_2, CO, O_3, and NO_2 at the five stations were determined to be highly accurate, accurate, and quite accurate, with Mean Absolute Percentage Error (MAPE) values rang- ing from 8% to 43%.
Keywords
Air Pollution, Pollutants, Air Quality Index (AQI), ARIMA, AQI Categories

Introduction

Air pollution arises from diverse sources, including industrial, transportation, and home sectors, exerting an influence on environmental well-being. Various variables, including as population expansion, high urbanization rates, inadequate spatial planning, and limited public awareness, indirectly contribute to air pollution (Simandjuntak, 2013). Air pollution is a matter of worldwide significance. Air pollution results in the introduction of dangerous solid, liquid, and gaseous chemicals into pure air, leading to contamination (Abidin and Hasibuan, 2019). The general public frequently underestimates the se- rious risk that polluted air poses to human health. Long-term, consistent rises in atmospheric pollution can result in

DOI: 10.6092/issn.2281-4485/20263

respiratory conditions such as emphysema, bronchitis, and lung cancer Sandhi, 2019. short term exposure to air pollution such as PM2.5 has also been linked to short term health problems such as cognitive decline (please enter reference). The mobility of the people in regions such as Jakarta might lead to an escalation in air pollution. Jakarta, the Special Capital Region of Indonesia, holds the distinction of being the most densely inhabited area in the country, with a population density of 15,978 individuals per square kilometer (km^2) Badan Pusat Statis tik, 2021. A high population density implies a widespread reliance on both private and public transportation for daily activities. In DKI Jakarta, there are a total of 21.7 million vehicles, with motorbikes being the most prevalent at 16.5 million units, followed by passenger cars, lorries, and buses Korlantas Kepolisian Republik Indonesia, 2022. DKI Jakarta frequently ranks as having the most unfavorable Air Quality Index (AQI) globally, as reported by IQAir (2023) Air, 2023. According to the UCAR Center for Science Education (2023), air quality refers to the measurement of pollutants present in a specific location, as shown by the air quality index for Science Education, 2023. Presently, Jakarta is equipped with a total of five air quality monitoring stations. There is a police post near the HI Roundabout in Central Jakarta (DKI1 station), one at the Kelapa Gading Subdistrict Office in North Jakarta (DKI2 station), one at the Taman Pendidikan Dinas Pertamanan in South Jakarta (DKI3 station), one on Pondok Gede Street in East Jakarta (DKI4 station), and one at Kebon Jeruk Residential Park in West Jakarta (DKI5 station). Jakarta claimed the top spot as the most polluted city in Indonesia in 2021, while Indonesia itself rated 17th among 118 nations with the most severe air pollution. In response to this, the Jakarta Environmental Agency devised an Air Pollution Control Strategy that aims to decrease the propor- tion of lethal pollution levels by 2030. This strategy functions as a point of reference in this research. Due to the substantial influence of air pollution on health, it is crucial to consistently regulate and oversee air pollution, in- cluding the use of prediction methods. This study focuses on forecasting the Air Quality Index (AQI) in Jakarta. Various statistical techniques, such as Autoregressive Integrated Moving Average (ARIMA), smoothing, regression, and econometrics, can be employed for prediction purposes (Box et al., 2015). The selection of the methodology is determined by the particular situation constraints, data patterns, preci- sion, and model observations. The choice of ARIMA in this work was based on its simplicity and ability to handle data with diverse patterns. ARIMA's three primary parameters, namely p, d, and q, reflect the autoregressive, differencing, and moving average components, respectively, providing flexibility in modeling. ARIMA can be applied to time series data with different properties by modifying these parameters. Furthermore, ARIMA can effectively address patterns and seasonal components. When seasonal patternsare detected in the data, the model will incorporate seasonal components to enhance the accuracy of forecasts and accommodate recurring changes. Seasonal data commonly displays patterns characterized by observations occurring at regular intervals of periods. Data visualization can be employed to ascertain seasonal values or components Dimashanti and Sugiman (2021). Additionally, the selection of ARIMA was based on a comparison analy- sis, which showed that the ARIMA (1,1,0) model outperforms the LSTM model with 7000 batches in forecasting CO levels Spyrou et al., 2022. In another study, it was discovered that the ARIMA model was more effective than the WNN and SVM models in forecasting air pollution. The ARIMA model achieved a R^2 value of 0.882, an MSE value of 0.056, and an NSE value of 0.880 Zhang et al., 2020. In terms of relevant study, utilizing the ARIMA (1,0,0) model to forecast the AQICO values in Surabaya. This model demonstrated the lowest Root Mean Square Error (RMSE) of 0.2349, indicating its superior performance Syaifulloh, 2021. In a separate research, an ARIMA(2,1,3)(1,0,0) model was employed to predict the levels of AQI PM_{10} in Bangalore Hosamane et al., 2020. This research aims to forecast the Air Quality Index (AQI) data for Jakarta using the Autoregressive Inte grated Moving Average (ARIMA) model, based on the provided information.

ARIMA Model

Autoregressive Integrated Moving Average (ARIMA), also known as the Box- Jenkins method, is a statistical technique used for the analysis and forecasting of time series data. The purpose of using this method in time series data analysis is to identify patterns and trends in the data and make predictions for future values. ARIMA demonstrates high forecasting accuracy for short- term or brief time periods but is typically less precise when used to forecast values over longer time frames because it tends to produce stable or constant estimates Adri Senen (2017).

DOI: 10.6092/issn.2281-4485/20263

ARIMA has three main components, namely AR (Autoregressive), I (Integrated), and MA (Moving Average) model. The AR model describes the relationship between the dependent variable (X) and its previous values. An AR model with order p is denoted as AR(p), expressed as follows:

$$\hat{X}_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + e_t \quad [1]$$

The MA model depicts the relationship between the dependent variable (X) and the previous values of residuals or errors. An MA model with order q is denoted as MA(q), expressed as follows:

$$\hat{X}_t = e_t - \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-p} + e_t \quad [2]$$

An ARIMA model with orders p, d, and q is denoted as ARIMA(p,d,q). In the ARIMA model, the differencing step (d) is applied to make the time series stationary before applying the ARMA components. For example, if $d = 1$ is performed, differentiating $(B)^1 X_t = W_t$ results in a horizontal pattern with an average of $\Phi = 0$, and the model is then written as an ARMA(p,q) model, expressed as follows:

$$\hat{W}_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + e_t \\ -\theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-p} + e_t \quad [3]$$

Methodology

This research is conducted using the Cross-Industry Standard Process for Data Mining (CRISP-DM) method. CRISP-DM is a method commonly used as an independent process by industries for data mining. The method stages can be viewed through the flowchart in Figure. 1. The first stage is business understanding, it involves defining the prob- lem, understanding business perspectives and needs, and planning to achieve goals in addressing air pollution in DKI Jakarta Province. Through fore- casting using the ARIMA method based on historical ISPU Jakarta data, information on air quality is obtained as a basis for government decision- making and public awareness of air health. This process contributes to pru- dent decision-making by authorities regarding air pollution and encourages public awareness to maintain air health. Additionally, literature review is conducted in this phase to serve as research references. The second is data understanding, it involves the process of understand- ing the data, encompassing data acquisition, description, exploration, and evaluation. The historical data used comprises the Air Quality Index (AQI) data for DKI Jakarta Province from 2010 to 2021, sourced from the public Jakarta Open Data platform. The variables utilized include the date and ISPU values for air pollutants PM_{10}, SO_2, CO, O_3, and NO_2. The following subsections outline the procedures involved in ARIMA model are data preparation, modeling, and evaluation Wei, 2018.

Pre-processing

Data integration, this stage is performed by consolidating all AQI data files at each station into one in '.csv' storage format. Data cleaning, this stage is carried out to remove unnecessary variables. During

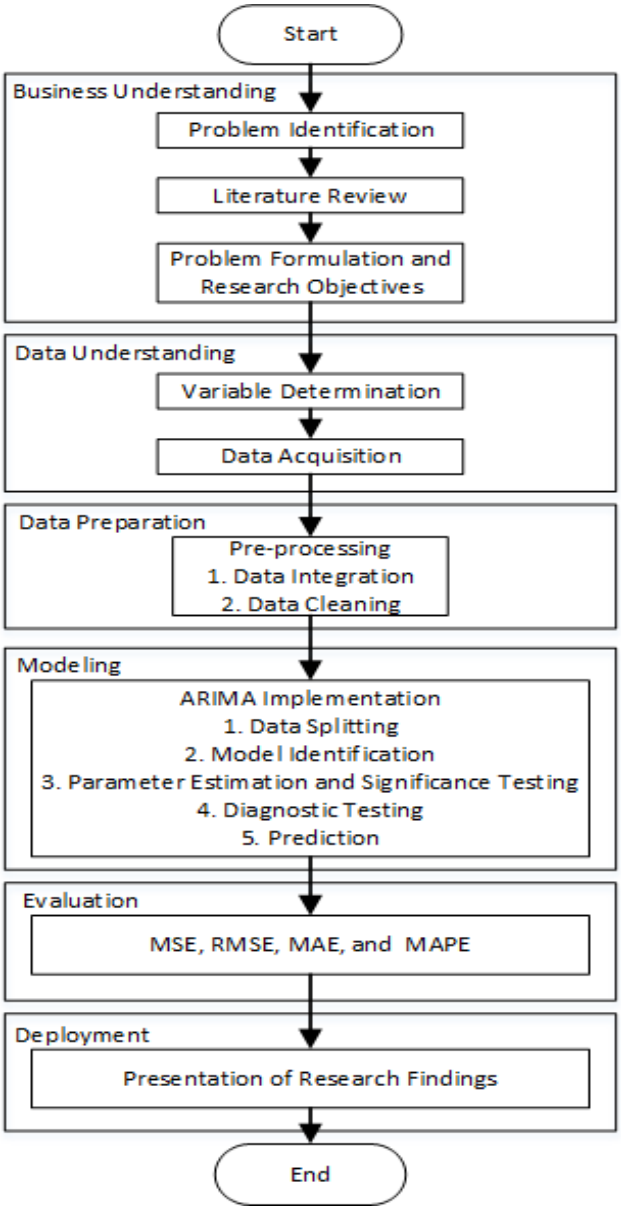

Figure 1. *Flowchart*

DOI: 10.6092/issn.2281-4485/20263

this phase, data with issues, such as missing data in a single row at the beginning of the year, will be dropped or removed entirely. For randomly missing data in specific cells, imputation will be performed by calculating replacement values using the average of values at X_{t-1} and X_{t-1} (where X_t represents the missing data).

Data Splitting

This step involves dividing the data into two types, namely, training data and test data. The test data consists of the last 7 data points for each pollutant at each station.

For PM_{10} pollutant, a total of 19,709 data were used for the training process and 35 data for testing. Specifically, DKI1 station utilized daily data from 2010 to 2021, with 4376 training data and 7 testing data (table 1). DKI2 station employed daily data from 2011 to 2021, with 4018 training data and 7 testing data. DKI3 station used daily data from 2011 to 2021, comprising 4018 training data and 7 testing data. DKI4 station employed daily data from 2011 to 2021, with 4018 training data and 7 testing data. DKI5 station utilized daily data from 2013 to 2021, with 3279 training data and 7 testing data.

Station	Time Frame	Training Data	Testing Data
DKI1	2010 - 2021	4376	7
DKI2	2011 - 2021	4018	7
DKI3	2011 - 2021	4018	7
DKI4	2011 - 2021	4018	7
DKI5	2013 - 2021	3279	7

Table 1: *Split the data for the PM_{10} pollutant*

For SO_2 pollutant, a total of 19,709 data were used for the training process and 35 data for testing. Specifically, DKI1 station utilized daily data from 2010 to 2021, with 4376 training data and 7 testing data. DKI2 station employed daily data from 2011 to 2021, with 4018 training data and 7 testing data (Table 2). DKI3 station used daily data from 2011 to 2021, comprising 4018 training data and 7 testing data. DKI4 station employ-yed daily data from 2011 to 2021, with 4018 training data and 7 testing data. DKI5 station utilized daily data from 2013 to 2021, with 3279 training data and 7 testing data.

Station	Time Frame	Training Data	Testing Data
DKI1	2010 - 2021	4376	7
DKI2	2011 - 2021	4018	7
DKI3	2011 - 2021	4018	7
DKI4	2011 - 2021	4018	7
DKI5	2013 - 2021	3279	7

Table 2. *Split the data for the SO_2 pollutant*

For *CO* pollutant, a total of 16,049 data were used for the training process and 35 data for testing. Specifically, DKI1 station utilized daily data from 2010 to 2021, with 4376 training data and 7 testing data. DKI2 station employed daily data from 2011 to 2021, with 4018 training data and 7 testing data (Table 3). DKI3 station used daily data from 2011 to 2021, comprising 4018 training data and 7 testing data. DKI4 station employed daily data from 2021, with 4018 training data and 7 testing data. DKI5 station utilized daily data from 2013 to 2021, with 3279 training data and 7 testing data.

Station	Time Frame	Training Data	Testing Data
DKI1	2010 - 2021	4376	7
DKI2	2011 - 2021	4018	7
DKI3	2011 - 2021	4018	7
DKI4	2021	358	7
DKI5	2013 - 2021	3279	7

Table 3. *Split the data for the CO pollutant*

For O_3 pollutant, a total of 1790 data were used for the training process and 35 data for testing. Specifically, DKI1 station utilized daily data from 2021, with 358 training data and 7 testing data. DKI2 station employed daily data from 2021, with 358 training data and 7 testing data (Table 4). DKI3 station used daily data from 2021, comprising 358 training data and 7 testing data. DKI4 station employed daily data from 2021, with 358 training data and 7 testing data. DKI5 station utilized daily data from 2021, with 358 training data and 7 testing data.

Station	Time Frame	Training Data	Testing Data
DKI1	2021	358	7
DKI2	2021	358	7
DKI3	2021	358	7
DKI4	2021	358	7
DKI5	2021	358	7

Table 4. *Split the data for the O_3 pollutant*

For NO_2 pollutant, a total of 19,709 data were used for the training process and 35 data for testing. Specifically, DKI1 station utilized daily data from 2010 to 2021, with 4376 training data and 7 testing data (Table 5). DKI2 station employed daily data from 2011 to 2021, with 4018 training data and 7 testing data. DKI3 station used daily data from 2011 to 2021, comprising 4018 training data and 7 testing da-

λValue -1 - 0.5 0 0.5 1

Transformation $1/X_t$ $\sqrt{1/X_t}$ λX_t $\sqrt{X_t}$ X_t

DOI: 10.6092/issn.2281-4485/20263

ta. DKI4 station employed daily data from 2011 to 2021, with 4018 training data and 7 testing data. DKI5 station utilized daily data from 2013 to 2021, with 3279 training data and 7 testing data.

Table 5. *Split the data for the NO_2 pollutant*

Station	Time Frame	Training Data	Testing Data
DKI1	2010 - 2021	4376	7
DKI2	2011 - 2021	4018	7
DKI3	2011 - 2021	4018	7
DKI4	2011 - 2021	4018	7
DKI5	2013 - 2021	3279	7

Modeling

Model identification. This is the stage of recognizing whether the data being used is stationary or not. In the context of ARIMA analysis, data must be stationary in both mean and variance. To determine if the time series data is stationary in mean, a test is conducted using the Dickey-Fuller test. If it turns out that the data is not stationary in mean, a differencing process of order d is performed. The following is the hypothesis statement for the Dickey-Fuller test:

1. H_0 = The data is not stationary in mean.
2. H_1 = The data is stationary in mean.
3. Determining $a = 0.05$
4. Critical area = H_0 is rejected if the p-value < a, meaning the data is stationary in mean.

Subsequently, to evaluate whether the data exhibits variance stationarity, a Box-Cox test is employed. In this context, the focus is on the value of λ. If $\lambda = 1$, it signifies non-stationarity in variance, necessitating a transformation process. To stabilize the variance using Box-Cox transformation with the parameter λ, as in this equation Heni Kusdarwati and Handoyo, (2018):

$$T(X_t) = X_t^{\lambda} = \frac{X_t^{\lambda} - 1}{\lambda} \qquad [4]$$

Various values of lambda and the transformation shapes resulting from equation [4] can be found in Table 6. The table illustrates various transformation functions applied to the variable X_t based on different λ values. These transformations play a crucial role in data manipulation and statistical modeling. The choice of transformation depends on the specific characteristics and distribution of the data under consideration. The next step is to identify the order of the ARIMA model based on the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. To estimate the values of p and q for the ARIMA model, one can observe the shapes of the ACF and PACF plots. Table 6 describes the patterns that may appear in the ACF and PACF plots in accordance with the ARIMA model theory. By referring to the ACF and PACF plots of the available data, several hypotheses can be proposed regarding the appropriate order values for ARIMA model development Heni Kusdarwati and Handoyo, 2018.

Table 6. *The form of transformation*

λValue	-1	- 0.5	0	0.5	1
Transformation	$1/X_t$	$\sqrt{1/X_t}$	λX_t	$\sqrt{X_t}$	X_t

The table 7 presents the ACF and PACF patterns for different time series models. The models considered include AR, MA, and ARMA. The ACF plot for an AR(p) model exhibits exponentially decreasing autocorrelation, while the PACF plot shows a cut-off at lag p. Conversely, for an MA(q) model, the ACF plot has a cut-off at lag q, and the PACF plot exhibits exponentially decreasing values. In the case of an ARMA(p,q) model, both the ACF and PACF plots demonstrate a tail-off pattern. Understanding these characteristic patterns is essential for model selection and parameter estimation in the analysis of time series data.

Table 7. *ACF and PACF Plot*

Model	ACF Plot	PACF Plot
AR(p)	Exponentially decreasing	Cut off at lag p
MA(q)	Cut off at lag q	Exponentially decreasing
ARMA(p,q)	Tail off	Tail off

Parameter estimation and significance testing. The next step is to seek estimates of the possible model values, followed by conducting tests to determine the significance of the parameters in the model. These tests are crucial to ensure that the chosen model has an adequate level of significance, indicating its suitability for use. In the Python programming language, model parameters are computed automatically using the Maximum Likelihood Estimation (MLE) method. Significance testing of the parameters is performed using the Wald test. The following is the hypothesis statement:

DOI: 10.6092/issn.2281-4485/20263

1. $H_0 : \beta_j = 0$ (The parameters are significant).
2. $H_1 : \beta_j$ 0, = 1, 2, 3, ..., k (The parameters are not significant).
3. Determining $a = 0.05$
4. Critical area = H_0 is rejected if the p-value < a, meaning the parame-ters are significant.

Diagnostic testing. After the identification and parameter estimation phase is completed, the next step is to perform diagnostic tests to evaluate the residuals of the model and test whether the model exhibits white noise and follows a normal distribution. Residuals are referred to as "white noise" if they do not meet the criteria for white noise Ahmar et al., 2018. The following is the hypothesis statement for the Ljung-Box test:
1. $H_0 : \varrho_j = 0$ (The residuals are white noise).
2. $H_1 : \varrho_j \neq 0$, = 1, 2, 3, ..., k (The residuals are not white noise).
3. Determining $a = 0.05$
4. Critical area = H_0 is rejected if the p-value < a, meaning the residuals do not meet the criteria for white noise.

The test to determine whether the residuals follow a normal distribution or not is conducted using the Shapiro-Wilk test. The following is the hypothesis statement for the Shapiro-Wilk test:
1. $H_0 : F(x) = F_0(x)$ (The residuals follow a normal distribution.).
2. $H_1 : F(x) \neq F_0(x)$ (The residuals do not follow a normal distribution.).
3. Determining $a = 0.05$
4. Critical area = H_0 is rejected if the p-value > a, meaning the residuals do not follow a normal distribution.

Prediction

The prediction process is performed by forecasting the training and test data using the selected tentative model.

Evaluation

Mean Squared Error (MSE). MSE is one of the error criteria in nonparametric regression. MSE is a metric that measures the average of the squared differences between the actual and predicted values Allen (1971). MSE penalizes larger errors more because each difference is squared. The formula is expressed in equation [5].

$$MSE = \frac{\sum_{t=1}^{n} \left(X_t - \hat{X}_t\right)^2}{n} \quad [5]$$

Root Mean Squared Error (RMSE). RMSE is the square root of MSE and has the same units as the target variable, which is typically a unit of time. It is used to measure how close the predicted values are to the actual values. Smaller RMSE indicates better model performance Jadon et al., 2022. The formula is expressed in equation [6]

$$RMSE = \sqrt{MSE} \quad [6]$$

Mean Absolute Error (MAE). MAE is a metric that calculates the average of the absolute differences between the predicted and actual values. Like MSE, MAE is used to measure the accuracy of a model, but it assigns equal weight to all errors, regardless of whether they are large or small Chicco et al., 2021. The formula is expressed in equation [7]:

$$MAE = \frac{\sum_{t=1}^{n} \left|X_t - \hat{X}_t\right|}{n} \quad [7]$$

Mean Absolute Percentage Error (MAPE). MAPE is a measure of accuracy obtained by calculating the distance be- tween actual and predicted data. MAPE is computed by taking the absolute errors within a specific time frame or season and dividing them by the actual data value, then expressing the result as a percentage De Myttenaere et al., 2016. The formula is expressed in equation [8]:

$$MAPE = \frac{\left(\frac{\sum_{t=1}^{n} \left|X_t - \hat{X}_t\right|}{X_t}\right)}{n} \times 100\% \quad [8]$$

According to Lewis (1982), MAPE consists of several levels of accuracy Lewis, 1982. The table 8 categorizes the MAPE into different levels of accuracy. MAPE is a widely used metric to assess the accuracy of forecasting models. These accuracy levels provide a standardized framework for interpreting MAPE results in various domains and applications.

MAPE (%)	Accuracy
< 10%	Highly accurate
10% – 20%	Accurate
20% – 50%	Quite accurate
> 50%	Inaccurate

Table 8 *The level of MAPE accuracy*

Results and Discussion

The following section presents an analysis of the ex-

DOI: 10.6092/issn.2281-4485/20263

perimental results for each pollutant at the five monitoring stations. The tables below present the results of modeling analysis, resulting in the best models with the lowest MAPE on the test data compared to other tentative models. In this comprehensive analysis, the performance of each pollutant model at the five monitoring stations is thoroughly examined to ensure the accuracy and reliability of the predictions. Furthermore, the identification of the best models with the lowest MAPE values on the test data is essential to provide policymakers and environmental authorities with valuable insights into air quality management strategies. Subsequently, the predicted AQI results are averaged for the year 2021 and categorized by pollutant. These aggregated values are then visualized using graphical plots, as depicted in the image below. These visual repre- sentations of AQI patterns serve as a vital resource for both researchers and the general public, promoting awareness and proactive measures to safeguard public health and the environment.

PM_{10} pollutant prediction results

The table 9 presents the ARIMA order and the corresponding MAPE for each monitoring station, providing insights into the accuracy of the PM_{10} predictions. The MAPE values indicate the accuracy of the PM_{10} predictions for each station. According to the predefined accuracy categories, DKI4's MAPE of 13.64% falls within the "Accurate" range, indicating a high level of precision in its forecasting. For DKI1, DKI2, DKI3, and DKI5, the MAPE values fall within the "Quite accurate" category, suggesting reliable forecasting performances.

Table 9 *ARIMA Model for PM_{10}*

Station	ARIMA Model	MAPE (%)
DKI1	AR(2)	20.79%
DKI2	AR(3)	21.75%
DKI3	AR(6)	21.17%
DKI4	AR(6)	13.64%
DKI5	AR(4)	22.20%

The above graph represents the Air Quality Index (AQI) values derived from the average predicted values of PM_{10} pollutant across 5 monitoring sta- tions in the year 2021 (Fig. 2). The graph illustrates that the PM_{10} AQI values are categorized as 'Good' at the beginning of the year, specifically in January,. Februaary, and March, as well as towards the end of the year in December. In the middle of the year, the average AQI values fall into the 'Moderate' category. The 'Good' category implies that PM_{10} has no significant adverse effects on living organisms, while the 'Moderate' category may result in re- duced visibility. The minimum recorded value is 30.34, while the maximum value reaches 72.47.

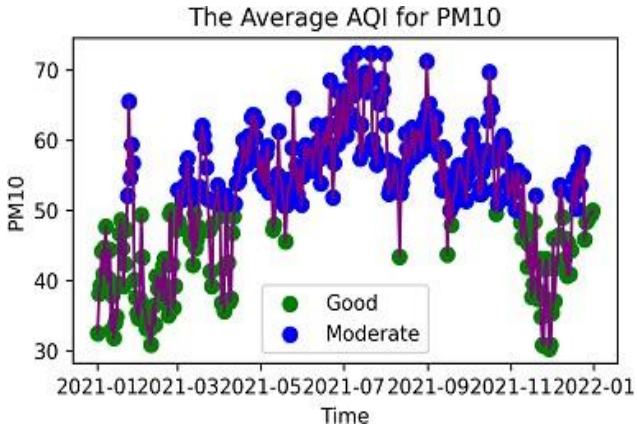

Figure 2. *The Average AQI for PM_{10} in Jakarta in 2021*

According to the Final Report on the Air Quality Monitoring in Jakarta (2020), the high values of PM_{10} pollutant Air Quality Index (AQI) are at- tributed to the elevated concentrations of AQI pollutants. This phenomenon is a consequence of the heavy traffic in Jakarta, where coarse particles on the road surface may be re-entrained into the air due to vehicular movement. The daily average of the PM_{10} pollutant parameter exhibits a strong correlation with various meteorological factors, as determined through correlation tests. It is positively correlated with temperature and radiation, while inversely related to humidity. Analyzing the concentration trends of PM_{10} during the COVID-19 pandemic reveals fluc- tuations with no significant differences. However, when compared to the average PM_{10} concentrations in 2020, which stood at 56.38 $\mu g/m^3$, there was a reduction. In previous years, the highest annual averages were 62.23 $\mu g/m^3$ in 2019 and 57.36 $\mu g/m^3$ in 2018 Jakarta, 2020a.

SO_2 pollutant prediction results

The table showcases the ARIMA order and corresponding MAPE for each monitoring station, providing insights into the accuracy of SO_2 predictions. The MAPE values indicate the precision of the SO_2 predictions for each station, with all stations falling within the "Accurate" range according to the predefined accuracy categories. Specifically,

DOI: 10.6092/issn.2281-4485/20263

DKI2's MAPE of 12.46%, DKI5's MAPE of 14.84%, and DKI1, DKI3, and DKI4 with MAPE values ranging from 15.13% to 17.86% all demonstrate accurate forecasting performances for SO_2 levels.

Station	ARIMA Model	MAPE (%)
DKI1	ARI(4,1)	15.13%
DKI2	ARI(4,1)	12.46%
DKI3	ARI(1,1)	17.86%
DKI4	ARI(4,1)	17.64%
DKI5	ARI(3,1)	14.84%

Table 10 *ARIMA Model for SO_2*

The graph above represents the AQI values derived from the average predicted values of SO_2 pollutant across five monitoring stations in the year 2021 (Fig. 3). The graph shows that the SO_2 AQI values tend to be high, categorized as 'Moderate,' and experience a significant increase in the middle of the year, indicated by the AQI value falling into the 'Unhealthy' category. The 'Good' and 'Moderate' categories signify that SO_2 can cause damage or injury to various types of plants when mixed with O_3 in a short period. The minimum recorded value is 21.14, and the maximum value reaches 117.47. The data visualization of SO_2 AQI demonstrates significant fluctuations. Based on meteorological factors, as indicated in the Final Report on Air Quality Monitoring in Jakarta (2020), the SO_2 pollutant parameter exhibits a direct correlation with temperature and radiation while showing an inverse relationship with humidity. The emissions sources of SO_2 are distributed across various sectors, with 61.96% originating from manufacturing industries, contributing approximately 2,637 tons/year. Additionally, 25.71% of the emissions come from the energy sector (around 1,071 tons/year),

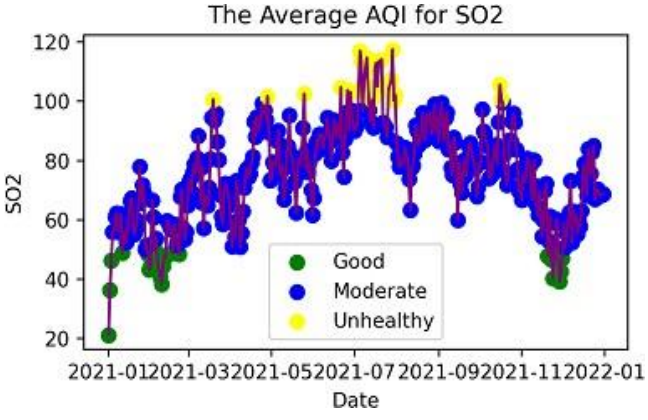

Figure 3. *The Average AQI for SO_2 in Jakarta in 2021*

11.58% from transportation (about 493 tons/year), and 1.26% from commercial and residential sectors (less than 42 tons/year) (Jakarta, 2020a). When considering the trends in SO_2 concentration during the COVID-19 pandemic, the annual average concentration of SO_2 in 2020, 25.14 $\mu g/m^3$, showed a decrease compared to the averages of 27.03 $\mu g/m^3$ in 2019, 20.05 $\mu g/m^3$ in 2018, and 26.24 $\mu g/m^3$ in 2017 Jakarta, 2020b. However, in the report records that a significant increase in SO_2 concentration is observed in 2021. This increase is attributed to the 11.01% growth in Gross Regional Domestic Product (GRDP) in 2021, as reported by the Jakarta Statistics Agency. The manufacturing industry, contributing 12.28%, was the second- largest contributor to the increase in SO_2 AQI values, following the large retail and trade sector. This growth in the manufacturing industry is likely responsible for the elevated SO_2 pollutant emissions Khoirunnisa (2023)..

CO pollutant prediction results

The table 11 presents the ARIMA order and corresponding MAPE for each monitoring station, providing insights into the accuracy of *CO* predictions. According to the predefined accuracy categories, DKI2's MAPE of 8.71% considered "Highly accurate," DKI1, DKI3, DKI4, and DKI5's MAPE, with MAPE values ranging from 16.19% to 43.39%, are labeled as "Quite accurate".

Station	ARIMA Model	MAPE (%)
DKI1	AR(5)	16.70%
DKI2	AR(6)	8.71%
DKI3	AR(6)	16.19%
DKI4	ARI(1,1)	43.39%
DKI5	AR(3)	19.98%

Table 11 *ARIMA Model for CO*

The graph above is a representation of the AQI values derived from the average predicted concentrations of CO pollutant at five monitoring stations in the year 2021 (Fig. 4) The graph shows that the CO AQI values consistently remain within the 'Good' category. The 'Good' category signifies that CO has no significant adverse effects on living organisms. The recorded values range from a minimum of 6.09 to a maximum of 41.57. The visualization of CO AQI data tends to experience an increase. From a meteorological perspective, based on the Final Report of Air Quality Monitoring in DKI Jakarta (2020), the CO

DOI: 10.6092/issn.2281-4485/20263

pollutant parameter exhibits an inverse correlation with radiation and temperature while showing a direct correlation with humidity. The report records a significant decrease in 2020, and it remains relatively stable with minimal fluctuations. This phenomenon can be attributed to government policies implemented during the COVID-19 pandemic, such as Large-Scale Social Restrictions. These policies led to a reduction in motor vehicle usage, resulting in decreased CO pollutant levels. When examining the average values, CO concentrations in 2017 were 1.49 $\mu g/m^3$, in 2018 were 1.40 $\mu g/m^3$, in 2019 were 1.49 $\mu g/m^3$, and in 2020 were 0.46 $\mu g/m^3$ Jakarta, 2020a.

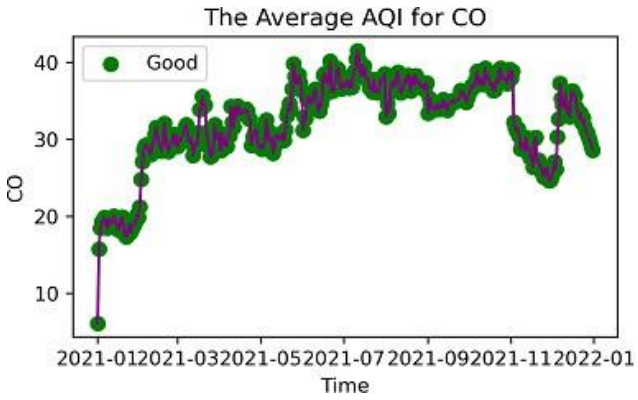

Figure 4: *The Average AQI for CO in Jakarta in 2021*

O_3 pollutant prediction results

The table presents the ARIMA order and the corresponding MAPE for each monitoring station, providing insights into the accuracy of the O_3 predictions. The MAPE values indicate the accuracy of the O_3 predictions for each station. According to the predefined accuracy categories, DKI3's MAPE of 19.12% and DKI3' MAPE of 19.72% falls within the "Accurate" range. For DKI1, DKI4, and DKI5, the MAPE values fall within the "Quite accurate" category, suggesting reliable forecasting performances. The graph above represents the AQI values derived from the average predicted concentrations of O_3 pollutant at five monitoring stations in the year 2021 (Fig. 5). The graph shows that the O_3 AQI values consistently remain within the 'Good' category. The 'Good' category implies that O_3 may cause damage to various types of plants when mixed with SO_2 for four consecutive hours. The recorded values range from a minimum of 6.99 to a maximum of 19.19. The visualization of O_3 AQI data exhibits a horizontal pattern. When examining the average values, O_3 concentrations in 2017 were 51.55 $\mu g/m^3$, in 2018 were 37.24 $\mu g/m^3$, in 2019 were 52.46 $\mu g/m^3$, and in 2020 were 56.78 $\mu g/m^3$. In 2021, there was a decrease in O_3 pollutant AQI values. This reduction was a result of the implementation of the Montreal Protocol by the United Nations, specifically the Kigali Amendment on the Reduction of Ozone-Depleting Substances (ODS) and Hydrofluorocarbon (HFC) to pro- tect the ozone layer, prevent extreme climate changes, and manage COVID- 19 vaccines using refrigeration systems. COVID-19 vaccines were managed from production, distribution, to safe and high-quality vaccination, all in accordance with the Montreal Protocol dan Kehutanan (BSILHK), 2021.

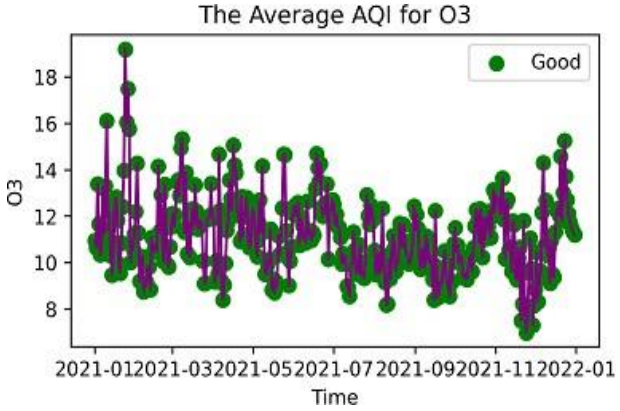

Figure 5. *The Average AQI for O_3 in Jakarta in 2021*

NO_2 pollutant prediction results

The table 13 presents the ARIMA order and corresponding MAPE for each monitoring station, providing insights into the accuracy of NO_2 predictions. According to the predefined accuracy categories, DKI2's MAPE of 9.28% and DKI1's MAPE of 10.15% considered "Highly accurate,"

Station	ARIMA Model	MAPE (%)
DKI1	AR(1)	21.07%
DKI2	AR(3)	19.72%
DKI3	AR(2)	19.12%
DKI4	AR(2)	28.50%
DKI5	AR(1)	28.43%

Table 12 *ARIMA Model for O_3*

Station	ARIMA Model	MAPE (%)
DKI1	AR(7)	10.15%
DKI2	AR(6)	9.28%
DKI3	AR(3)	15.59%
DKI4	AR(6)	30.75%
DKI5	AR(4)	26.84%

Table 13 *ARIMA Model for NO_2*

DOI: 10.6092/issn.2281-4485/20263

DKI3's MAPE of 15.59% is labeled "Accurate. DKI4 and DKI5 with MAPE values are 26.84% and 30.75%, are labeled as "Quite accurate". The graph below represents the AQI values derived from the average predicted concentrations of NO_2 pollutant at five monitoring stations in the year 2021 (Fig. 6). The graph shows that the NO_2 AQI values consistently remain within the 'Good' category. The 'Good' category implies that NO_2 may produce a slight odor. The recorded values range from a minimum of 6.49 to a maximum of 51.07. The visualization of NO_2 displays a horizontal pattern and tended to increase at the beginning of the year. When considering the meteorological factors, based on the Final Report of Air Quality Monitoring in DKI Jakarta (2020), the correlation between NO_2 pollutant parameters is inversely related to radiation and temperature, and directly related to humidity. The report records significant differences in the years 2016, 2017, and 2018, which had lower AQI values compared to the following years. In terms of average values, NO2 concentrations in 2017 were 13.51 $\mu g/m^3$, in 2018 were 10.08 $\mu g/m^3$, in 2019 were 46.6 $\mu g/m^3$, and in 2020 were 34.50 $\mu g/m^3$ Jakarta, 2020a.

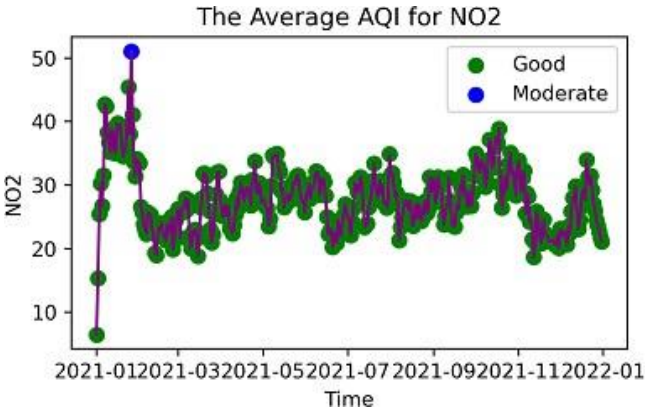

Figure 6. *The Average AQI for NO_2 in Jakarta in 2021*

Conclusions

Predicting the AQI of PM_{10}, SO_2, CO, O_3, and NO_2 pollutants at the monitoring stations DKI1, DKI2, DKI3, DKI4, and DKI5 in Jakarta using ARIMA resulted in 25 models with highly accurate, accurate, and reasonably accurate performance, each with a Mean Absolute Percentage Error (MAPE) ranging from 8% to 43% on the test data. This could provide guidance for the government in the decision-making process regarding air pollution control in Jakarta, and for the public to consistently maintain air quality for their health.

References

ABIDIN J., HASIBUAN F.A. (2019) Pengaruh dampak pencemaran udara terhadap kesehatan untuk menambah pemahaman masyarakat awam tentang bahaya dari polusi udara. Prosiding Seminar Nasional Fisika Universitas Riau IV (SNFUR-4) Pekanbaru, 7 September 2019, 4(2):3. ISBN: 978-979-792-691-5

AHMAR A.S., GURITNO S., RAHMAN A., MINGGI I., TIRO M.A., AIDID M.K., ANNAS S., SUTIKSNO D.U., AHMAR D.S., AHMAR K.H. ET AL.(2018) Modeling data containing outliers using ARIMA additive outlier (ARIMA-AO). Journal of Physics,Conference Series, 954(1):012010 https://doi.org/10.1088/1742-6596/954/1/012010

AIR I.Q. (2023) Kualitas udara jakarta (tech. rep.). Index Quality Air. https://www.iqair.com/id/indonesia/jakarta

ALLEN D.M. (1971) Mean square error of prediction as a criterion for selecting variables. Technometrics, 13(3): 469–475. https://doi.org/10.2307/1267161

BPS - Badan Pusat Statistik. (2021) Kepadatan Penduduk DKI Jakarta (tech. rep.). Badan Pusat Statistik.

CHICCO D., WARRENS M.J., JURMAN G. (2021) The coefficient of deter- mination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. PeerJ Computer Science, 7:e623.

DAN KEHUTANAN B.S.I.L.H. (2021) Hari ozon sedunia tahun 2021: Peran rantai pendingin di masa pandemi (tech. rep.). Kemente-rian Lingkungan Hidup dan Kehutanan.

DE MYTTENAERE A., GOLDEN B., LE GRAND B., ROSSI F. (2016) Mean absolute percentage error for re-gression models. Neurocomputing, 192:38–48.

DIMASHANTI A.R., SUGIMAN (2021) Peramalan indeks harga konsumen kota semarang menggunakan sarima berbantuan software minitab. PRISMA, Prosiding Seminar Nasional Matematika, 4:565–576.

HENI KUSDARWATI U.E., HANDOYO S. (2018). Analisis deret waktu univariat linier. John Wiley & Sons.

HOSAMANE S.N., PRASHANTH K., VIRUPAKSHI A. S. (2020) Assessment and prediction of PM10 concen-tration using ARIMA. Journal of Physics: Conference Se-ries, 1706(1):012132. https://doi.org/10.1088/1742-6596/1706/1/012132

JADON A., PATIL A., JADON S. (2022) A comprehensive survey of regression based loss functions for time series forecasting. Machine Learning, Cornell University, pp.13. https://doi.org/10.48550/arXiv.2211.02989

JAKARTA D.L.H.D. (2020a) Laporan akhir (januari-desember) pemantauan kualitas udara dki jakarta tahun 2020 (tech. rep.). Dinas Lingkungan Hidup DKI Jakarta.

DOI: 10.6092/issn.2281-4485/20263

JAKARTA D.L.H.D. (2020b) Laporan inventarisasi emisi pencemar udara dki jakarta tahun 2020 (tech. rep.). Dinas Lingkungan Hidup DKI Jakarta.

KHOIRUNNISA (2023) Industri manufaktur di Jakarta alami pertumbuhan (tech. rep.). Dinas Lingkungan Hidup DKI Jakarta.

KKRI - Korlantas Kepolisian Republik Indonesia. (2022). Jumlah Kendaraan Bermo- tor Menurut Jenis Kendaraan (Unit) di Provinsi DKI Jakrta (tech. rep.). Kepolisian Republik Indonesia.

LEWIS C.D. (1982) Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting. Butterworth Scientific, 143 pp. ISBN 0408005599, 9780408005593

SANDHI S.I. (2019) Studi fenomenologi: Kesadaran diri (self awareness) per- okok aktif yang mempunyai anak balita dalam perilaku merokok di tempat umum di kelurahan pegulon kabupaten kendal. Jurnal Ke- bidanan Harapan Ibu Pekalongan, 6:237–243.

SENEN A., RATNASARI T. (2017) Studi peramalan beban ratarata jangka pendek menggunakan metoda autoregressive integrated moving average (ARIMA). Jurnal Ilmiah Sutet, 7(2):93–101. https://doi.org/10.33322/ sutet.v7i2.84

SIMANDJUNTAK A.G. (2007) Pencemaran udara. Buletin Limbah, 11(1).

SPYROU E.D., TSOULOS I., STYLIOS C. (2022) Applying and comparing LSTM and ARIMA to predict CO levels for a time-series measurements in a port area. Signals, 3(2):235–248. https://doi.org/10.3390/signals3020015

SYAIFULLOH M.M. (2021) Prediksi indeks standar pencemaran udara di kota surabaya berdasarkan konsentrasi gas karbon monoksida. Jambura Journal of Probability and Statistics, 2(2):86–95.

UCAR (2024) What is air quality? (Tech. rep.). Center for Science Education. https://scied.ucar.edu/learning-zone/air-quality/what-is-air-quality

WEI W.W. (2019) Multivariate time series analysis and applications. John Wiley & Sons. ISBN: 978-1-119-50285-2

ZHANG Y., YANG H., CUI H., CHEN Q. (2020) Comparison of the ability of ARIMA, WNN and SVM Models for drought forecasting in the Sanjiang Plain, China. Natural Resources Research, 29(6). https://doi.org/10.1007/s11053-019-09512-6