Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)

Journal Homepage: http://journal.lembagakita.org/index.php/jtik

Forecasting the Air Quality Index by Utilizing Several Meteorological Factors Using the ARIMAX Method (Case Study: Central Jakarta City)

Naufal Fadli Muzakki [1], Azmi Zulfani Putri [2], Surya Maruli [3], Fitri Kartiasih [4*]

[1,2,3] Statistical Computing Study Program, Politeknik Statistika STIS, East Jakarta City, Special Capital Region of Jakarta, Indonesia.

[4] Statistics Study Program, Politeknik Statistika STIS, East Jakarta City, Special Capital Region of Jakarta, Indonesia.*

a r t i c l e i n f o

Article history:
Received 1 January 2024
Received in revised form
4 March 2024
Accepted 20 April 2024
Available online July 2024.

DOI:
https://doi.org/10.35870/jtik.v8i3.2012.

Keywords:
Air Quality Index;
Meteorology; ARIMAX;
Central Jakarta.

Kata Kunci:
Indeks Kualitas Udara;
Meteorologi; ARIMAX;
Jakarta Pusat.

a b s t r a c t

Today's society really pays attention to air quality because the impact of exposure to pollutants in the air is starting to be felt. PM 2.5 pollutants are very dangerous because their small size can penetrate the alveoli of human lungs. The value calculation of the Air Quality Index (AQI) is important to prepare mitigation and defensive measures to reduce the negative impact of air quality and as a basis for future policymaking. Several method comparisons have been carried out by researchers to predict AQI. However, researchers have not studied much regarding the use of meteorological factors in the form of average air temperature (°C), average air humidity (percent), and average wind speed (m/s) in forecasting AQI values, even though meteorological factors have a significant link, according to previous researchers. This research forecasts AQI using the ARIMAX method, which includes meteorological factors as exogenous variables, using daily AQI PM 2.5 data in Central Jakarta. The best modeling of the data is ARIMA (1,1,1) without X and ARIMAX (1,1,1). Based on the calculation of AIC, BIC, RMSE, and MAPE values, ARIMAX (1,1,1) modeling produces better forecasting, so it can be concluded that forecasting involving meteorological factors can make forecasting more precise. Predicting AQI using ARIMAX with upcoming meteorological factors is beneficial, as precise prediction results can assist in policy-making to prevent the adverse impacts of air quality on public health. In future research, other meteorological factors could be studied and combined with other modeling besides ARIMA.

a b s t r a k

Masyarakat saat ini sangat memperhatikan kualitas udara karena dampak paparan polutan di udara mulai terasa. Polutan PM 2.5 sangat berbahaya karena ukurannya yang kecil dapat menembus alveoli paru-paru manusia. Penghitungan nilai Indeks Kualitas Udara (AQI) penting untuk mempersiapkan langkah mitigasi dan defensif guna mengurangi dampak negatif kualitas udara dan sebagai dasar pengambilan kebijakan di masa depan. Beberapa perbandingan metode telah dilakukan peneliti untuk memprediksi AQI. Namun peneliti belum banyak mengkaji mengenai penggunaan faktor meteorologi berupa suhu udara rata-rata (°C), kelembaban udara rata-rata (persen), dan kecepatan angin rata-rata (m/s) dalam meramalkan nilai AQI, padahal faktor meteorologi memiliki hubungan yang signifikan, menurut peneliti sebelumnya. Penelitian ini meramalkan AQI dengan menggunakan metode ARIMAX yang memasukkan faktor meteorologi sebagai variabel eksogen dengan menggunakan data harian AQI PM 2.5 di Jakarta Pusat. Pemodelan data terbaik adalah ARIMA (1,1,1) tanpa X dan ARIMAX (1,1,1). Berdasarkan perhitungan nilai AIC, BIC, RMSE, dan MAPE, pemodelan ARIMAX (1,1,1) menghasilkan peramalan yang lebih baik, sehingga dapat disimpulkan bahwa peramalan yang melibatkan faktor meteorologi dapat membuat peramalan menjadi lebih tepat. Memprediksi AQI menggunakan ARIMAX dengan faktor meteorologi yang akan datang sangatlah bermanfaat, karena hasil prediksi yang tepat dapat membantu dalam pengambilan kebijakan untuk mencegah dampak buruk kualitas udara terhadap kesehatan masyarakat. Pada penelitian selanjutnya, faktor meteorologi lain dapat dipelajari dan dikombinasikan dengan pemodelan lain selain ARIMA.

Corresponding Author. Email: fkartiasih@stis.ac.id [4].

ACM Computing Classification System (CCS)
EBSCOhost
Communication and Mass Media Complete (CMMC)

© E-ISSN: 2580-1643.

1. Introduction

Humans need air to breathe, so air is an important factor supporting human life. Over the last few decades, researchers have proven that poor air quality can encourage respiratory diseases that cause death [1]. This research is supported by evidence reports from the World Health Organization (WHO) that exposure to pollutants and poor air quality can cause cardiovascular mortality and morbidity, which can result in a decrease in acute respiratory diseases in future generations [2]. Air quality is getting worse due to exposure to particulate matter (PM) and other harmful substances resulting from industrialization and urbanization [3]-[5]. In recent decades, public attention about the importance of healthy air has increased. Especially in the capital city of Jakarta, people are starting to feel the many impacts of bad air. It is necessary to provide daily data and accurate forecast data so that the public has an idea of air quality conditions.

Air Quality Index (AQI) and PM 2.5 are two important indicators among pollution indices. AQI is a measurement metric used to standardize values and understand air quality and pollutant levels in it [4]. In Indonesia, AQI is calculated using a pollutant concentration measurement tool. AQI is calculated with reference to the new ambient air quality standard (GB3095-2012), which covers six pollutants, including sulfur dioxide (SO2), nitrogen dioxide (NO2), PM 2.5, PM 10, ozone (O3), and carbon monoxide (CO). One of the pollutants calculated is PM 2.5. PM 2.5 pollutant is dangerous for humans because its size is very small, namely below 2.5 μm. The human body is unable to filter small particles; as a result, they can enter the alveoli in the lungs and even penetrate the lungs and then exit into the bloodstream, causing significant health problems [7]. Daily AQI 2.5 data in Central Jakarta shows the average change in trend as the day changes. For the general public, AQI is an important index for understanding whether air quality is good or bad. Communities need AQI values to take defensive steps so as not to be affected by bad air quality.

Poor air quality with high levels of PM2.5 pollutants in the long term can affect human health [8], so estimates of air quality are needed to determine appropriate policies in the future. Several researchers have studied AQI (Air Quality Index) forecasting using many methods. Mathematically, in air quality forecasting, there are two types of models used: deterministic models and empirical models. Deterministic models, such as chemical transport models (CTMs), apply basic principles to simulate the atmospheric chemical and physical processes involved in the emission, transport, and transformation of air pollution. However, due to the complexity of chemical and transport processes in the atmosphere, these models have significant estimation and uncertainty, leading to a lack of accuracy compared to empirical models of air quality that have been developed, fine-tuned for specific locations, and trained with local meteorological data [9]. In the context of empirical models, various statistical models are used for air quality indices. For example, there are models such as ARIMA (Autoregressive Integrated Moving Average), Multiple Linear Regression (MLR), Artificial Neural Networks (ANN), Support Vector Regression (SVR), and also a hybrid model.

The ARIMA model is a classic technique in statistical analysis that is used to understand non-linear time series data [10]. The ARIMA model, which is a combination of autoregression and moving average, utilizes previous values in a time series to make projections about the future. The ARIMA approach is considered useful because it is able to reflect the temporary persistence of air pollutants in the atmosphere in the model, so ARIMA is a powerful and useful statistical tool in evaluating and forecasting air pollution levels [11]. In previous research, the use of the ARIMA model for forecasting was usually combined with other models such as ARIMA-LTSM [12] or hybrid ARIMA [13] to increase the Air Quality Index's (AQI) forecasting precision. However, there is still very little research that examines the use of the ARIMAX model (Autoregressive Integrated Moving Average with Explanatory Variable) to perform AQI prediction.

According to several studies, meteorological factors include minimum air temperature (°C), maximum air temperature (°C), average air temperature (°C), average air humidity (percent), duration of sunlight (hours), speed Maximum wind (m/s) and average wind speed (m/s) significantly influence Air Quality [14]. This is also supported by studies that

unfavorable meteorological factors can influence the formation and growth of new air pollutants as well as the ability of the atmosphere to disperse air pollutants. Meteorological factors are significant predictors in estimating the concentration of submicron particles in the atmosphere which are used as indicators for AQI measurements [15]. This shows that forecasting AQI by considering meteorological factors is a good thing to do, several previous studies have also suggested considering other factors when forecasting. The aim of this research is to provide an overview of air quality conditions, build an air quality forecasting model and explore the effectiveness of meteorological factors which are exogenous variables in predicting air quality in Central Jakarta.

In order to take effective action as a preventive measure to avoid the impact of poor air quality, it is important to forecast using other exogenous variable approaches that influence AQI to be able to understand AQI better. Therefore, this research focuses on describing air quality conditions, building an air quality forecasting model using ARIMA(X), and comparing the precision of air quality forecasting results with the ARIMA and ARIMA(X) models. The data used is daily AQI PM 2.5 data for the last 2 years and seven meteorological factor data in Central Jakarta, which is one of the most critical and important areas in the capital city of Jakarta with quite poor air quality.

This research includes describing and forecasting air quality conditions in Central Jakarta, which takes into account exogenous variables in the form of meteorological factors that have not been widely studied. First, this research answers several previous studies that found that meteorological factors simultaneously influence air quality with the right combination of variables. Most previous research, such as Wood [16] using data mining; Liu & Guo [17] using the LTSM method; and Miri et al. [18] using a regression method, only tried one combination of meteorological variables. This research will compare several combinations of meteorological variables to find the combination of meteorological factors that most significantly influences air quality in Central Jakarta. Second, this research predicts AQI with a non-linear method using exogenous variables, which, according to several researchers, is more precise than deterministic methods. Not many studies have studied AQI forecasting using the ARIMAX method; previous research has mostly studied AQI forecasting using ARIMA or deterministic methods such as [19]–[21]. Third, this research also looks at whether forecasting with ARIMAX provides more precise results than regular ARIMA. This research answers the assumption from previous research that imprecision in AQI forecasting is possible due to the influence of meteorological factors in the atmosphere [22]. This research is expected to provide benefits and become a reference for future forecasting involving other meteorological factors to provide more precise forecasting.

Air Quality Index (AQI)

According to reference [23], the Air Quality Index (AQI) is a metric used to assess the level of pollution in the air in a given area. An essential tool for accurately assessing the state of the air in a given location is the Air Quality Index (AQI). The air quality index is computed using a range of pollutants, including carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), fine particulate matter (PM2.5), particulate matter (PM10), and ground-level ozone (O3) [24] . Greater values of the index indicate poorer air quality and higher levels of pollution in the area. On the other hand, a lower index indicates lower pollution levels and improved air quality in the area [25].

Fine particulate matter 2.5 (PM2.5) is defined as air particles with a diameter of less than 2.5 microns [26]. Dust and engineering projects—which typically involve bacteria, oxidants, minerals, and other elements—are the main sources of these airborne particles. Some of them are also legitimate fine particles because they are produced by natural processes such sandstorms, volcanic eruptions, forest fires, and waves [27]. Burning garbage and fossil fuels like coal and oil produces derivatives of sulfur and nitrogen oxides that are then converted into fine particulate matter (PM2.5). Burning fossil fuels also releases greenhouse gases into the atmosphere. As a result of their continued reliance on coal and oil as energy sources, developing nations like Indonesia are typically the ones affected. Additionally, the technology in use lacks adequate waste gas processing, and other energy sources have not been created [27]. The oxidation of other exhaust gas pollutants and

direct emissions from vehicles are two major sources of fine particulate matter (PM2.5), as is tire and brake deterioration. Another important source of fine particulate matter (PM2.5), particularly in the summer, is combustion from heating and culinary activities like burning wood. Indoor activities including cooking, smoking, and candle lighting can also produce PM2.5 emissions [28]. In certain environments, indoor PM2.5 concentrations can reach hundreds of μg m3. Due to its high liquidity, fine particulate matter (PM2.5) during the rainy season travels quickly to nearby areas and urban areas when air flow is affected, having an impact across a larger area [27].

Human premature death from respiratory and cardiovascular disorders is closely correlated with both fine particulate matter (PM2.5) and particulate matter (PM10) [29]. According to a number of studies, tiny particulate matter (PM2.5) poses a greater risk than fine particulate matter (PM10). Fine particulate matter (PM2.5) is the fifth most deadly risk factor for death and is projected to cause 4.2 million deaths annually worldwide [30]. According to reference [31], fine particulate matter (PM2.5) is the primary risk factor for respiratory diseases in China, accounting for over 1.1 million premature deaths from conditions like stroke, ischemic heart disease (CHD), lung cancer, lower respiratory tract infections, and severe chronic obstructive pulmonary disease (COPD).

Meteorology and Air Pollution

Air pollutants are gases or airborne particulates that have diverse chemical and physical compositions, causing a range of impacts on air quality that are harmful to human health [32]. It is believed that weather patterns and air pollution are related, or vice versa. Numerous investigations have been conducted to examine the connection between air pollution and weather patterns [33]. The minimum, maximum, average, and average air temperatures as well as the amount of sunshine, wind speed, and humidity are all considered meteorological parameters.

The lowest temperature attained in a certain place or area is referred to as the minimum air temperature. When temperatures are often at their lowest throughout the day, such as in the morning, minimum air temperatures are frequently recorded. The highest temperature measured over a given duration, typically six, twelve, or twenty-four hours, is referred to as the maximum air temperature. Due to the heat continuing to build after noon, the air temperature reaches its peak about 15.00 on average. The highest temperature of the day is typically measured once a day around nine in the morning local time. The term "average air temperature" describes the mean temperature of the atmosphere in a given area for a given time frame, like a day, month, or year. Water vapor in the air divided by the greatest quantity that the air can contain at a given temperature is the measure of average air humidity. Different times and places have different average humidity levels.

Observer latitude and season have an impact on how long sunshine lasts. The sun's course determines how long sunlight lasts, while the distance from the equator affects how long sunsets last. Air movement across a surface in a horizontal direction is known as wind speed. This velocity can be expressed as either an instantaneous speed (called wind gusts) or an average velocity over a 2-minute period. By collecting samples and averaging the data, the maximum wind speed can be found. samples collected throughout a specific time frame. Mean wind speed, which is commonly expressed in miles per hour (mph) or meters per second (m/s), is the average wind speed for a specific time period.

2. Research Methods

This research uses secondary data obtained from the World Air Quality Project Historical Air Quality Data Platform (https://aqicn.org) and the BMKG Online Database Center. The data used is daily individual AQI data on PM2.5 (μm/m3), minimum air temperature (°C), maximum air temperature (°C), average air temperature (°C), average air humidity (percent), duration of sunlight (hours), maximum wind speed (m/s), and average wind speed (m/s) from November 1, 2021, to October 31, 2023, so there are 760 observations. PM2.5 individual AQI is measured using the Beta Attenuation Monitoring method. Average air temperature, average air humidity, and average wind speed were collected from the Kemayoran Meteorological Station. This research explores the effectiveness of several meteorological factors as exogenous variables in predicting the daily

air quality index. The stages that must be carried out in this research are explained in Figure 1. Air Quality Index data from the World Air Quality Project is combined with meteorological data from BMKG into one data set. Then it's done by inputting the missing value using k-nearest neighbors to make the data clean for use in model building. After obtaining clean data, model formation and model evaluation were carried out to obtain the best model for forecasting the Central Jakarta City Air Quality Index.

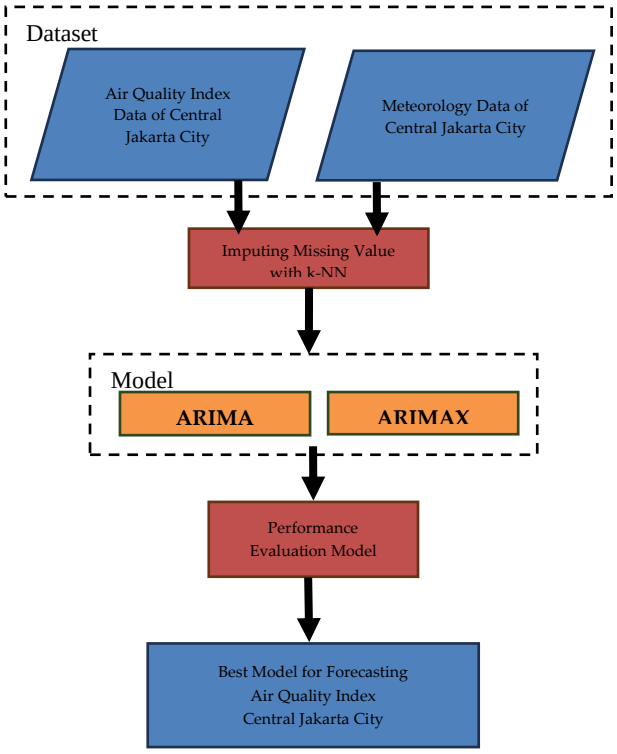

Figure 1. Stages of forming a prediction model for the Daily Air Quality Index for Central Jakarta City

In calculating the Air Quality Index, various pollutant indices (CO, NO2, SO2, PM2.5, PM10, O3, NH3, and Pb) are used for calculations. The Air Quality Historical Data Platform (the World Air Quality Project) only provides Air Quality Index data, which is measured based on the PM2.5 and PM10 pollutant indices. The data set contains a number of missing data values. The availability of the Air Quality Index calculated based on the PM2.5 pollutant index is better than other pollutant indices. Therefore, the PM2.5 pollutant index is the main data for predicting the Central Jakarta City Air Quality Index. The meteorological variables provided by BMKG consist of many variables, such as minimum air temperature, maximum air temperature, average air temperature, average humidity, rainfall, duration of sunlight, maximum wind speed, average wind speed, and current wind direction. maximum speed, and maximum wind direction. Reference [14] stated that meteorological factors consisting of rainfall, intensity of solar radiation, wind speed, air humidity, and air temperature influence the level of air pollution. Another study by reference [34] stated that wind speed, air temperature, air pressure, and relative humidity influence the level of air pollution. The rainfall variable data provided by BMKG contains many unmeasurable values. Therefore, the variables used as exogenous variables for predicting the Central Jakarta City Air Quality Index in this study are minimum air temperature, maximum air temperature, average air temperature, average air humidity, duration of sunlight, maximum wind speed, and average wind speed.

After selecting variables and obtaining clean data, it was discovered that in the data set there were several missing values. The completeness of observation data is an important requirement that must be met in advanced analyses such as time series analysis [35]. There are many ways to deal with missing values. In this study, k-NN (k-nearest neighbor) was used to fill in missing values (imputing missing values). k-NN can replace missing data by using the average value of the k nearest neighbors. In the context of the imputation of missing data in a time series, k-NN can produce accurate estimates by considering the values before and after the missing data. The k-NN method was chosen because k-NN has been proven to provide the best results in reconstructing missing data and makes a positive contribution to time series forecasting compared to other imputation methods [35]-[36]. The results of four experimental cases show that the k-NN technique is most effective in reconstructing missing data and makes a positive contribution to time series forecasting compared to other imputation methods. For example, on the CNNpred dataset with the lowest loss rate (1.86%), the results show that k-NN widens the performance gap compared with other imputation methods. Therefore, these results indicate that the k-NN method has a positive impact on time series forecasting analysis and provides better results than other imputation methods. Thus, it can be concluded that the k-NN imputation results positively influence the time series forecasting analysis.

There is a limitation to imputation using the k-NN method, namely that this method requires calculating the distance between missing data points and existing data points. Therefore, if there are many dimensions or variables in the dataset, calculating these distances can become complex and time-consuming. Additionally, k-NN is also susceptible to outliers and noise in the data, which can affect the imputation results. Potential bias in imputation using the k-NN method can arise if there are not enough relevant or representative data points in the vicinity of the missing data points. This can lead to inaccurate imputation and affect the performance of the forecasting model. However, in this study, the limitations and potential bias of imputation using k-NN can be avoided because the data used in this study is daily data on air quality and meteorology, which have close inter-temporal characteristics, so that the appearance of relevant or representative data around the data points is assessed. What is missing is enough representation.

In this study, data was imputed using the k-NN method with a k value of 5. The data set that was imputed using k-NN is shown in Figure 2.

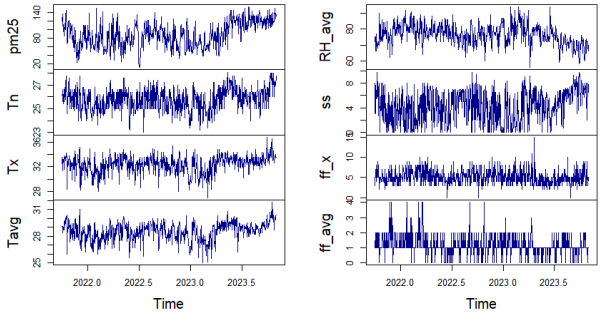

Figure 2. Graph of the data set after imputing using k-NN

The ARIMA (Autoregressive Integrated Moving Average) model is the most commonly used method in time series analysis. ARIMA combines the concepts of autoregression and moving average to provide a prediction value formed from a linear combination of previous variable values and prediction errors [35]. ARIMA consists of three parameters: the autoregressive order p represents the number of lags of Y that will be used as a predictor, the integration order d states the number of differentiation times to make the data stationary, and q represents the moving average order, which represents the number of forecast errors in the past period that must be included in the ARIMA model [38]. In general, ARIMA can be described using the following mathematical model.

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Where t is the observation value at time t, p is the parameter of the autoregressive component of the model, q is the moving average component parameter, and t is the error in the t-th period [36].

Not much different, the ARIMAX model is a development of the ARIMA model. This model develops the ARIMA model, whose predictions are not only influenced by a linear combination of previous variable values and prediction errors but are also influenced by exogenous variables marked with the letter X [37]. The inclusion of exogenous variables that are proven to influence the predicted value can improve forecasting ability [39]. In general, ARIMAX can be described using the following mathematical model.

$$Y_t = c + \beta_1 X_{1,t} + \cdots + \beta_s X_{s,t} + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where t is the observation value at time t, s is the parameter for the independent variable Xs, p is the parameter of the autoregressive component of the model, q is the moving average component parameter, and t is the error in the t-th period [37].

The exogenous variables that will be used in building the model to help predict the air quality index are several meteorological variables in the form of minimum air temperature (Tn), maximum air temperature (Tx), average air temperature (Tavg), average air humidity (RH_avg), duration of sunlight (ss), maximum wind speed (ff_x) and average wind speed (ff_avg).

In the ARIMA model, the selection needs to be made by selecting the best model among the possible ARIMA models that exist. The model lag level (p,d,q) can be selected based on the Akaike information

criterion (AIC), corrected Akaike information criterion (AICc), or Bayesian information criterion (BIC) values [40]. The mathematical model of these three criteria is as follows:

$$AIC = -2\ log\ log\ (L)\ + 2k,$$
$$AICc = AIC + 2k(k+1)n - k - 1,$$
$$BIC = -2log(L) + klog(n),$$

where L is the likelihood function of the data in the model, n is the sample size, and k are the total parameters in the model, with k=p+q+d+1 if the model intercept is zero, and k=p+q+d if the model intercept is not zero [41].

In this study, Root Mean Square Error (RMSE) and Mean Average Percentage Error (MAPE) were calculated to assess the forecasting accuracy of each model. RMSE is a measure to measure the difference between the values predicted by a time series model and the actual data. RMSE can be written as the following equation.

$$RMSE = \sqrt{\frac{\sum \varepsilon_t^2}{n}}$$

By forecasting error at the t and nth times, many observation time periods are predicted [42]. MAPE is another frequently used measure, calculating the average of the percentage differences between predicted values and actual values. MAPE can be written as the following equation.

$$MAPE = \frac{\sum \frac{|\varepsilon_t|}{X_t}}{n}$$

where t forecasting error at time t, n many observation time periods are forecast, and Xt is the actual value at time t [42].

To get the best ARIMA model, there are several assumptions that must be met on the residuals, namely White Noise, normal distribution test, residual average test, and arch test. The White Noise Test can be used using the Ljung-Box test. The Normal Distribution Assumption Test is carried out to determine whether the residuals are normally distributed or not. The Shapiro-Wilk Test is not used. The average residual test uses the t test to determine whether the residual is around 0. Meanwhile, the arch test is carried out to test whether the residual is homogeneous.

3. Result and Discussion

The following is an overview of the Central Jakarta City Air Quality Index from November 1, 2021, to October 31, 2023, to answer the first objective of this research, and it also presents a scatter plot matrix of several variables used in predicting the air quality index. Figure 3 is a heatmap diagram that shows the level of air pollution based on the Air Quality Index. The diagram divides air pollution levels into six levels, as shown in Table 1.

Table 1. Air Pollution Levels Based on the Air Quality Index

Air Quality Index ($\mu m/m^3$)	Air Pollution Level	Color
0 – 50	Good	Green
51 – 100	Moderate	Yellow
100 – 150	Unhealthy for Sensitive Groups	Orange
151 – 200	Unhealty	Red
201 – 300	Very Unhealty	Purple
300+	Hazardous	Maroon

Source : The World Air Quality Project (https://aqicn.org)

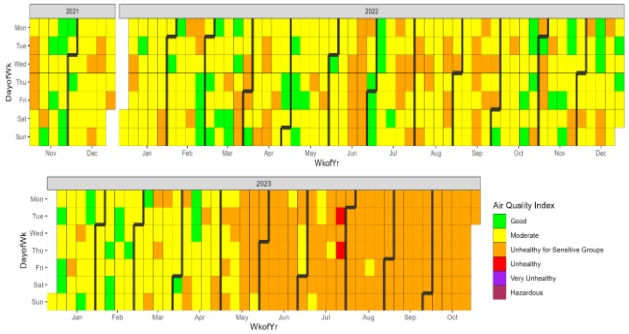

Figure 3. Air Pollution Levels based on the Central Jakarta City Air Quality Index

The worst air pollution levels occurred in the last 5 months (June–October) in 2023, where the average air pollution level was categorized as "unhealthy for sensitive groups" to "unhealthy" in that time period. Meanwhile, for the previous period, the average level

of air pollution was at the "moderate" level. In general, air pollution conditions in Central Jakarta City are rarely at the "good" level.

Figure 4. Correlation between variables used in research

The relationship between the Air Quality Index variable (PM25) and the duration of sunlight (ss) is positive, with a Pearson correlation value of 0.23. This value indicates a weak relationship between the two variables. Then, minimum air temperature (Tn) has a positive relationship with the air quality index (PM25) with a Pearson correlation value of 0.35. Judging from the Pearson correlation value, the relationship between the two variables tends to be weak. Maximum air temperature (Tx) has a positive relationship with the air quality index (PM25) with a Pearson correlation value of 0.24, which means there is a weak relationship between the two variables. Average air temperature (Tavg) also has a positive and weak relationship with the air quality index (PM25), namely the Pearson correlation value of 0.34. Meanwhile, the average air humidity variable (RH_avg) has a negative and weak relationship with the air quality index variable (PM25) with a Pearson correlation value of -0.26. The maximum wind speed variable (ff_x) also has a weak and negative relationship with the air quality index variable (pm25), where the Pearson correlation value is -0.17. Not unlike the average wind speed variable (ff_avg), it also has a Pearson correlation value of -0.30, which can be interpreted as having a weak and negative relationship with the air quality index (pm25).

A positive relationship can be interpreted as every increase in the duration of sunlight (ss), minimum air temperature (Tn), maximum air temperature (Tx), and average air temperature (Tavg), which will be followed by an increase in the air quality index (PM25). Meanwhile, a negative relationship can be interpreted as meaning that every increase in average air humidity (RH_avg), maximum wind speed (ff_x), and average wind speed (ff_avg) will be followed by a decrease in the air quality index (pm25). The results of the correlation exploration of each exogenous variable that will be used in forming the model show a weak relationship with the Air Quality Index variable. This occurs because the correlation analysis here is carried out partially for each variable. Meanwhile, the model formation is carried out simultaneously, so the simultaneous combination of exogenous variables is expected to help predict the Air Quality Index.

To avoid spurious regression, in modeling using time series analysis, it is necessary to pay attention to whether the data is stationary. Data is stationary when it is around the mean and variance. In this research, to prove that the data is stationary or not, a formal test was carried out using the Augmented Dickey Fuller (ADF) test with $\alpha = 0.05$. If the p-value > α ADF test conditions are met, then the null hypothesis cannot be rejected, and this means the data is not stationary. The ADF test results using R software can be seen in Table 2.

Table 2. Augmented Dickey Fuller (ADF) test results

Variable	Dickey-Fuller Statistics	p-value
Air Quality Index	-5.4439	0.01
minimum air temperature	-5.6499	0.01
maximum air temperature	-5.7172	0.01
average air temperature	-5.4107	0.01
long exposure to sunlight	-5.6047	0.01
average air humidity	-6.3985	0.01
maximum wind speed	-6.9464	0.01
average wind speed	-7.4962	0.01

Source: The World Air Quality Project and BMKG (2021-2023), data processed

Then this data is used as input for the modeling process to obtain the best model. After the data is stationary, the next process for selecting and optimizing ARIMA parameters is to observe the ACF/PACF graph and EACF matrix. Once this is achieved, the selected models are compared based on their AIC and BIC values. The best-fit model will be the one with the lowest AIC and BIC values. The auto.arima () function available in R software is used in this research, this function carries out the checking process steps as explained previously. When the function produces a result, the residuals of the resulting model are checked. If it resembles white noise, the forecast can then be calculated with optimal values for the parameters p, d, and q.

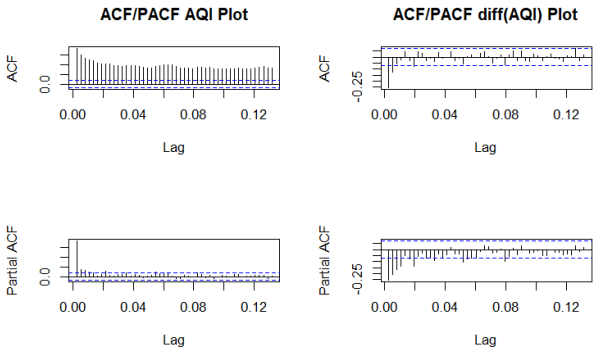

Figure 5. ACF and PACF plots for Air Quality Index data

As a starting point, we chose to generate a model using the auto.arima() function. The models returned by auto.arima() have the values p = 2, d = 1, and q = 2, respectively. To verify whether these models are the best fit or not, we apply the manual ARIMA function by trying various possible parameters for p, d, and q following the steps described previously.

Table 3. Performance Comparison of Tentative ARIMA Models

Model	AIC	BIC	Parameter Significance
Auto ARIMA(2,1,2)	6369.046	6396.596	AR(2) is not significant
ARIMA(1,0,0)	6434.344	6448.123	Significant
ARIMA(1,0,1)	6413.188	6431.561	Significant
ARIMA(2,0,1)	6378.125	6401.09	Significant
ARIMA(2,0,2)	6379.406	6406.965	MA(2) is not significant
ARIMA(1,1,1)	6366.212	6379.987	Significant
ARIMA(2,1,1)	6367.532	6385.899	AR(2) is not significant
ARIMA(1,1,2)	6367.419	6385.786	MA(2) is not significant

Source: The World Air Quality Project and BMKG (2021-2023), data processed

The ARIMA model obtained from the auto.arima () function is not necessarily the best result, this can be seen in Table 3. Although the AIC and BIC values are relatively small compared to the others, the AR (2) parameter is not significant in the ARIMA (2,1,2). As a comparison, it was found from manual experiments that the parameters ARIMA (1,1,1) produced AIC and BIC values that were even smaller than ARIMA (2,1,2) with all significant parameters in the model. Therefore, after comparing the measurement results, we conclude that the ARIMA (1,1,1) model is the most suitable model for the Air Quality Index data. The ARIMA (1,1,1) air quality index forecasting model obtained from R software is written as follows.

$$Y_t = c + 0{,}533508Y_{t-1} - 0{,}955229\varepsilon_{t-1} + \varepsilon_t$$

where Yt is the air quality index in period t, Yt-1 is the air quality index in 1 period before t, ε_(t-1) is the error in 1 period before t and ε_t is the error in period t.

As a comparison, modeling is carried out using ARIMAX, where exogenous variables will be added that have been proven to influence the Air Quality Index in previous research. The exogenous variables used in this research are meteorological variables in the form of minimum air temperature (Tn), maximum air temperature (Tx), average air temperature (Tavg), average air humidity (RH_avg), duration of sunlight

(ss), maximum wind speed (ff_x), and average wind speed (ff_avg). It is known that for some meteorological variables, there are several types of measurements, such as minimum, maximum, and average. So, this research will be carried out through trial and error for each possible model. On software R, there is a function ols_step_best_subset(), which is also used in this research to find the best combination of exogenous variables to include in the model. The ARIMA parameters used are the same as the best results in previous modeling, namely p = 1, d = 1, and q = 1. The results of testing the ARIMAX(1,1,1) model for each subset of tentative exogenous variables are presented in the following table.

Table 4. Tentative model selection ARIMAX (1,1,1)

Exogenous Variables	AIC	BIC	Parameter Significance ($\alpha = 0{,}05$)
Tavg, Tn, RH_avg, ff_avg	6313.181	6345.323	Tn is not significant
Tavg, ss, RH_avg, ff_avg	6310.952	6343.094	ss is not significant
Tx, ss, RH_avg, ff_avg	6350.415	6382.557	Tx and ss
Tn, ss, RH_avg, ff_avg	6342.723	6374.864	not significant
Tavg, ss, RH_avg, ff_x	6320.024	6352.166	significant
Tx, ss, RH_avg, ff_x	6359.365	6359.365	ss and ff_x
Tn, ss, RH_avg, ff_x	6350.893	6383.034	not significant
Tavg, RH_avg, ff_avg	6311.822	6339.372	Tx, ss and ff_x
Tx, RH_avg, ff_avg	6349.936	6377.486	not significant
Ss, RH_avg, ff_avg	6350.8	6378.35	ss and ff_x
RH_avg dan ff_avg	6351.037	6373.995	not significant
Tavg, ss, RH_avg	6318.208	6345.758	significant
Tavg, RH_avg, ff_x	6320.448	6347.998	Tx is not significant
Tavg dan RH_avg	6318.502	6341.46	ss is not significant
ss, RH_avg, ff_x	6359.336	6386.886	significant
RH_avg dan ff_x	6359.232	6382.19	ss is not significant
Ss dan RH_avg	6357.427	6380.385	ff_x is not significant
Tx, RH_avg, ff_x	6358.689	6386.239	significant
Tx dan RH_avg	6356.818	6379.777	ss and ff_x
Tx, ss, RH_avg	6357.592	6385.142	not significant
RH_avg	6357.249	6375.616	ff_x
Tn, RH_avg, ff_x	6352.5	6380.05	not significant
Tn, ss, RH_avg	6348.973	6376.523	ss
Tn, RH_avg	6350.503	6373.461	not significant

Source: The World Air Quality Project and BMKG (2021-2023), data processed

The exogenous variable subset obtained from the ols_step_best_subset() function is not necessarily the best result in model building, this can be seen in Table 5, although the AIC and BIC values are relatively small compared to the others, the Tn variable is not significant. For comparison, it was found from manual experiments that the subset of variables Tavg, RH_avg, ff_avg produced AIC and BIC values in the ARIMAX(1,1,1) model which were even smaller than the previous subset of exogenous variables with all significant variables in the model. Therefore, after comparing the measurement results, we conclude that the ARIMAX(1,1,1) model with exogenous variables average temperature (Tavg),

average humidity (RH_avg), and average wind speed (ff_avg) is a model with a subset of exogenous variables that best fits the Air Quality Index data and its supporting variables. The ARIMAX (1,1,1) air quality index forecasting model obtained from R software is written as follows.

$$Y_t = 8{,}8\,T_{avg_t} + 1{,}503\,RH_{avg_t} - 3{,}072 ff_{avg_t} + 0{,}532\,Y_{t-1} - 0{,}953\varepsilon_{t-1} + \varepsilon_t$$

where Y_t is the observation value in period t, T_{avg_t} is the average air temperature in period t, RH_{avg_t} is the average air humidity in period t, ff_{avg_t} is the average wind speed in period t, Y_{t-1} is the observation value in 1 period before t, ε_{t-1} is the error in 1 period before t and ε_t is the error in period t.

To test the normality of errors in the ARIMAX(1,1,1) model, the Shapiro-Wilk test and Q-Q plot visualization were carried out using R software. Residual pattern testing in this study also used R software, which was seen from the residual plot and supported by the t test using the hypothesis that zero on average is 0. Then, to test whether the residuals are correlated, the Ljung-Box test is carried out and also observed from the ACF/PACF plot using R software. In addition, it is necessary to confirm whether data volatility needs to be modeled using the Breusch-Pagan Test on R software. A summary of the output is shown in the following figure and table.

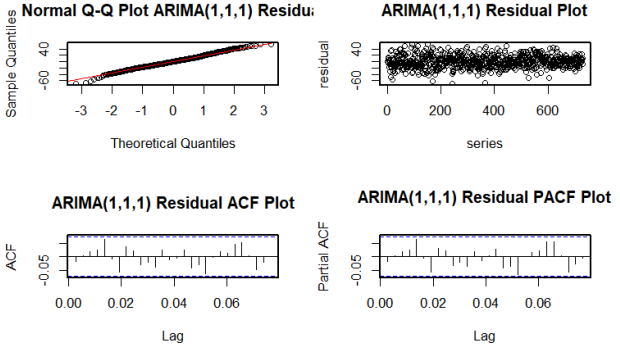

Figure 6. ARIMA (1,1,1) Diagnostic Test using Plot

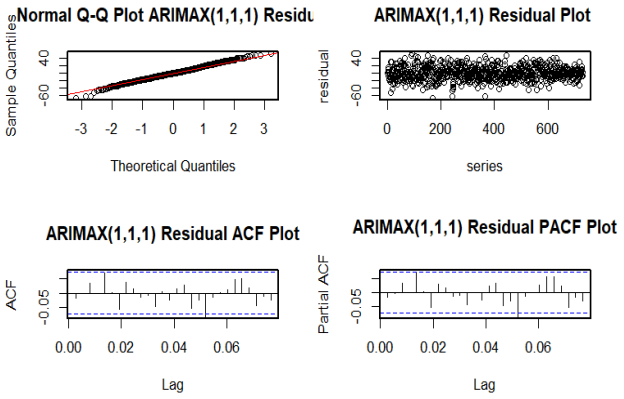

Figure 7. ARIMAX (1,1,1) Diagnostic Test using Plot

Table 5. Diagnostic Test Results for ARIMA (1,1,1) and ARIMAX (1,1,1) models using Formal Tests

Test Type	p-value	
	ARIMA (1,1,1)	ARIMAX (1,1,1)
Shapiro-Wilk Normality Test	0.02815	0.105
Box-Ljung Test	0.6427	0.657
One Sample t-test	0.4032	0.3945
ARCH LM-test	0.005323	0.005035

Based on the Q-Q plot graphs in Figures 6 and 7, it can be seen that the residuals for the two distribution models do not follow the normal line, but in this visualization, it is not clear that normality is visible. error, so a formal test was carried out using the Shapiro-Wilk test. In the Shapiro Wilk test results, the ARIMA model (1,1,1) shows a p-value of 0.02815. Because the p-value is smaller than alpha 5 percent, the decision to reject H0 is given, where H0 indicates the error is normally distributed and H1 indicates the error is not normally distributed. Thus, it can be concluded that with a significance level of 5 percent, there is not enough evidence to state that the ARIMA (1,1,1) model error follows a normal distribution. In contrast to the ARIMAX(1,1,1) model, where the p-value is greater than alpha 5 percent, it gives a decision to reject H0, which means that with a significance level of 5 percent, there is sufficient evidence to state that the ARIMAX(1,1,1) model error follows a normal distribution. From these results, it is known that, in terms of fulfilling the normal distribution assumption, ARIMAX (1,1,1) is considered better.

In the ACF and PACF plots, the observed values of the ARIMA (1,1,1) model do not appear to cross the significant line, or it can be said that the residuals are independent of each other. However, for the ARIMAX (1,1,1) model there is lag which exceeds the Bartlett line. This needs to be tested further with the results of the formal Ljung-Box test obtainedp-value for the ARIMA (1,1,1) model of 0.6427 > α = 0.05, then it fails to reject H0, where H0 indicates error not correlated and H1 showserror correlated. That is, there is not enough evidence to state that the remainder is intermediate lag mutually correlated, or it can be said that there is no autocorrelation between the residuals lag at a real level of 5%. Likewise with the ARIMAX (1,1,1) model, where the valuep-value is 0.657 greater than alpha 5 percent, then it gives a failed decision to reject H0, which means that with a significance level of 5 percent, there is not enough evidence to state that the ARIMAX (1,1,1) models are correlated with each other lag.

The residual plots of ARIMA (1,1,1) and ARIMAX (1,1,1) show the same pattern the residuals are random and scattered around the zero value. This is also supported by the results of the one-sample t-test where values were obtained p-value for the ARIMA (1,1,1) model is 0.4032 > α = 0.05 and for the ARIMAX (1,1,1) model, the p-value = 0.3945 > α = 0.05, then it fails to reject H0, where H0 shows the residual is around the value 0 and H1 vice versa. This means that there is sufficient evidence to state that the mean value of the residuals for both models is equal to zero at the 5% level of significance.

Based on the test results using the Breusch-Pagan Test, a value was obtained p-value for the residual (error) of the ARIMA (1,1,1) model is 0.005323 and the ARIMAX (1,1,1) model is 0.005035. Because p-value is smaller than 5 percent, it results in a decision to reject H0, where H0 shows a homogeneous residual variety and H1 shows an inhomogeneous residual variety. This means that with a significance level of 5 percent there is not enough evidence to say error for both homogeneous models.

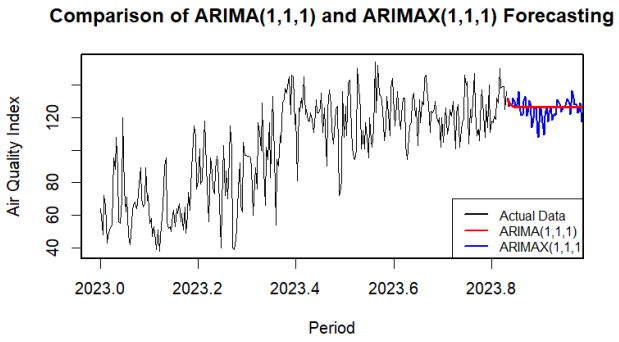

Figure 8. A comparative graph of ARIMA and ARIMAX Air Quality Index forecasting

After obtaining the best model to predict the Air Quality Index, in this research, a forecast of the air quality index was carried out for November 2023. It can be seen that the forecasting results using the ARIMA (1,1,1) model provide forecasts that tend to be constant and do not follow the actual data pattern of the previous period. The situation is different from the results of forecasting using the ARIMAX (1,1,1) model with exogenous variables like average air temperature, average air humidity, and average wind speed, which provide forecasts that tend to follow the actual data pattern of the previous period. The forecast values for the Central Jakarta City Air Quality Index using the two models are presented in figures 9 and table 6 below.

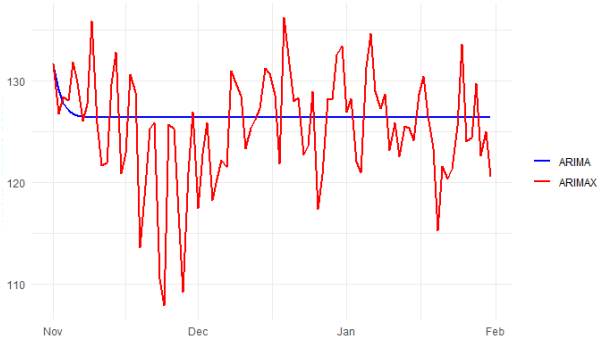

Figure 9. Air Quality Index forecasting graph

Table 6. AQI forecasting results for January 2024

Tanggal	ARIMA (1,1,1)	ARIMAX (1,1,1)
01/01/2024	126.39	126.98
02/01/2024	126.39	128.27
03/01/2024	126.39	122.04
04/01/2024	126.39	120.96

05/01/2024	126.39	131.38
06/01/2024	126.39	134.71
07/01/2024	126.39	128.99
08/01/2024	126.39	127.34
09/01/2024	126.39	128.67
10/01/2024	126.39	123.14
11/01/2024	126.39	125.91
12/01/2024	126.39	122.59
13/01/2024	126.39	125.55
14/01/2024	126.39	125.40
15/01/2024	126.39	124.15
16/01/2024	126.39	128.75
17/01/2024	126.39	130.46
18/01/2024	126.39	126.58
19/01/2024	126.39	123.36
20/01/2024	126.39	115.24
21/01/2024	126.39	121.55
22/01/2024	126.39	120.35
23/01/2024	126.39	121.28
24/01/2024	126.39	125.56
25/01/2024	126.39	133.61
26/01/2024	126.39	124.04
27/01/2024	126.39	124.38
28/01/2024	126.39	129.73
29/01/2024	126.39	122.62
30/01/2024	126.39	125.01
31/01/2024	126.39	120.50

Forecasting results using the best model, ARIMAX(1,1,1) show fluctuating data patterns according to past data patterns. A decrease or increase in AQI values can be caused by the influence of exogenous variables such as average air temperature, average air humidity, and average wind speed. For example, an increase in average air temperature and average wind speed can cause a decrease in AQI values because warmer air conditions and faster winds can help in the dispersion of air pollutants. Conversely, an increase in average air humidity can cause an increase in AQI values because high air humidity can worsen air quality by extending the residence time of pollutants in the air. This is in line with the correlation analysis in Figure 4 as well as previous research on the relationship between air quality and meteorological factors by reference [12]. Thus, adding exogenous variables to the ARIMAX (1,1,1) model can help predict changes in AQI values more accurately, because these variables can provide additional information about meteorological conditions that affect air quality.

To determine the effectiveness of adding exogenous variables like average air temperature, average air humidity, and average wind speed to the ARIMA model, the values were calculated in the error model in the form of RMSE and MAPE for each model. The results of calculating the RMSE and MAPE values are presented in the following table.

Table 7. Performance comparison of ARIMA (1,1,1) and ARIMAX (1,1,1) models

Jenis	ARIMA (1,1,1)	ARIMAX (1,1,1)
RMSE	18.94876	18.1804
MAPE	21.04613	20.19012
AIC	6366.212	6311.822
BIC	6379.987	6339.372

The RMSE and MAPE values for the ARIMAX (1,1,1) model are relatively smaller compared to the RMSE and MAPE for the ARIMA (1,1,1) model. This means that the ARIMAX (1,1,1) model has better forecasting performance than the ARIMA (1,1,1) model. This is supported by the AIC and BIC values which also show smaller values for the ARIMAX (1,1,1) model than for the ARIMA (1,1,1) model.

The results of this research have important implications in relation to the environment and public health. This research highlights the detrimental impacts of poor air quality, especially related to fine particulate PM2.5, which can cause serious respiratory and cardiovascular health illnesses. The implication of this research is the importance of understanding the relationship between meteorological factors and air quality to take appropriate mitigation measures. By considering factors such as air temperature, air humidity, and wind speed in predicting the air quality index (AQI), this research shows that the use of the ARIMAX model can increase the accuracy of AQI predictions. This can help governments and public health authorities adopt more effective policies to protect public health from the impacts of adverse air quality.

In addition, this research also highlights the importance of accurate meteorological monitoring in predicting air quality. By understanding the relationship between meteorological factors and air pollution, governments and related agencies can develop more effective strategies for managing air quality and protecting public health. Overall, this research makes an important contribution to understanding the relationship between air quality, meteorological factors, and public health. The implications of this research can help in the development of more effective environmental and public health policies to protect society from the impacts of adverse air quality.

4. Conclusions

In recent decades, there has been an issue with poor air quality. Respiratory conditions like cardiovascular disease can be brought on by poor air quality. One indicator that can be used to gauge the quality of the air in a given location is the Air Quality Index (AQI). One of the pollutants used to calculate the Air Quality Index (AQI) is fine particulate matter (PM2.5). Any airborne particle less than 2.5 micrometers in diameter is referred to as fine particulate matter, or PM2.5. Data from the Historical Air Quality Data Platform (The World Air Quality Project) and the BMKG Online Database Center were utilized in this study. The daily individual AQI PM2.5 data (μm/m3), minimum and maximum air temperatures (°C), average and minimum air temperatures (°C), average air humidity (percent), hours of sunlight, and average and maximum wind speeds (m/s) were obtained. There were missing values discovered during the preparation of the data. Using a k value of 5, we applied the k-NN approach to fill in the missing values in order to overcome them.

With exogenous factors including average air temperature, average air humidity, and average wind speed, this study generates ARIMA (1, 1, 1) and ARIMAX (1, 1, 1) models to predict the air quality index. In contrast Based on RMSE, MAPE, AIC, and BIC values, one may compare the accuracy of the ARIMA (1,1,1) model with the ARIMAX (1,1,1) model. The RMSE value in the ARIMAX (1,1,1) model comes out to be 18.1804. where the RMSE value is 18.94876 and the MAPE value is 21.04613 in the ARIMA (1,1,1) model, and the MAPE value is 20.19012 in that model. Based on these findings, the ARIMAX (1, 1, 1) model outperforms the ARIMA (1, 1, 1) model in terms of forecasting. This is supported by the AIC (6311.822) and BIC (6339.372) values in the ARIMAX (1,1,1) model, which are smaller than the AIC (6366.212) and BIC (6379.987) in the ARIMA (1,1,1) model.

The findings of the best model's forecasting, ARIMAX (1, 1, 1), reveal varying data patterns based on historical data patterns. The AQI number can be impacted by exogenous factors such wind speed, air temperature, and air humidity. Air quality index (AQI) can be increased or decreased depending on the temperature and wind speed. By include more details about the climatic factors that impact air quality, these variables can increase the accuracy of AQI forecasts made using the ARIMAX (1, 1, 1) model. This study's shortcomings include the fact that it only used data from Central Jakarta City, making it unable to forecast the air quality index over a larger region. Additionally,

only missing value detection and handling are done in this study; outlier data are not taken into account. The ARIMAX model may be impacted by outlier data in a number of ways, including altered autocorrelation and partial autocorrelation patterns, altered model parameter values, and elevated forecasting error values.

Subsequent studies might employ other techniques to complement the ARIMAX forecasting model with alternative forecasting models that might be more accurate and appropriate for air quality index forecasting. To create a more accurate forecasting model, it is also essential to pay attention to the data utilized to identify and eliminate missing values and outliers. It is anticipated that future studies would take into account additional exogenous variables in order to improve the accuracy of PM2.5 concentration forecasts. In order to produce better data that can aid in the prediction of air quality indices, it is desired that BMKG can concentrate on enhancing the precision and accuracy of meteorological monitoring. The way the government carries out its plans to control the quality of the air in Central Jakarta needs to be improved. Starting with the implementation of meteorologically responsive air management policies—such as restricting transportation or industrial activity when conditions have the potential to deteriorate air quality—this can be accomplished.

5. References

[1] Jhun, I., Coull, B. A., Schwartz, J., Hubbell, B., & Koutrakis, P. (2015). The impact of weather changes on air quality and health in the United States in 1994–2012. *Environmental research letters*, *10*(8), 084009.

[2] Ren, C., & Tong, S. (2008). Health effects of ambient air pollution–recent research development and contemporary methodological challenges. *Environmental Health*, *7*, 1-10.

[3] Qin, S., Liu, F., Wang, C., Song, Y., & Qu, J. (2015). Spatial-temporal analysis and projection of extreme particulate matter (PM10 and PM2. 5) levels using association rules: A case study of the Jing-Jin-Ji region, China. *Atmospheric Environment*, *120*, 339-350. DOI: https://doi.org/10.1016/j.atmosenv.2015.09.006.

[4] Kartiasih, F., & Setiawan, A. (2020). Aplikasi error correction mechanism dalam analisis dampak pertumbuhan ekonomi, konsumsi energi dan perdagangan internasional terhadap emisi CO2 di Indonesia. *Media Statistika*, *13*(1), 104-115. DOI: https://doi.org/10.14710/medstat.13.1.104-115.

[5] Kartiasih, F., & Pribadi, W. (2020). Environmental quality and poverty assessment in Indonesia. *Jurnal Pengelolaan Sumberdaya Alam Dan Lingkungan (Journal of Natural Resources and Environmental Management)*, *10*(1), 89-97. DOI: https://doi.org/10.29244/jpsl.10.1.89-97.

[6] Hageneder, S., Bauch, M., & Dostalek, J. (2016). Plasmonically amplified bioassay–Total internal reflection fluorescence vs. epifluorescence geometry. *Talanta*, *156*, 225-231.

[7] Hu, D., & Jiang, J. (2014). PM 2.5 pollution and risk for lung cancer: a rising issue in China. *Journal of Environmental Protection*, *2014*.

[8] Doherty, R. M., Heal, M. R., & O'Connor, F. M. (2017). Climate change impacts on human health over Europe through its effect on air quality. *Environmental Health*, *16*, 33-44.

[9] Cobourn, W. G. (2010). An enhanced PM2. 5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. *Atmospheric Environment*, *44*(25), 3015-3023. DOI: https://doi.org/10.1016/j.atmosenv.2010.05.009.

[10] Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., & Che, J. (2017). Daily air quality index forecasting with hybrid models: A case in China. *Environmental pollution*, *231*, 1232-1244. DOI: https://doi.org/10.1016/j.envpol.2017.08.069.

[11] Hsu, H. H., Adamkiewicz, G., Houseman, E. A., Spengler, J. D., & Levy, J. I. (2014). Using mobile monitoring to characterize roadway and aircraft contributions to ultrafine particle concentrations near a mid-sized airport. *Atmospheric Environment*, *89*, 688-695. DOI: https://doi.org/10.1016/j.atmosenv.2014.02.023.

[12] Wen, T., Liu, Y., he Bai, Y., & Liu, H. (2023). Modeling and forecasting CO2 emissions in China and its regions using a novel ARIMA-LSTM model. *Heliyon*, *9*(11).

[13] Zhao, L., Li, Z., & Qu, L. (2022). Forecasting of Beijing PM2. 5 with a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition. *Heliyon*, *8*(12).

[14] Aprianto, Y., Nurhasanah, N., & Sanubary, I. (2018). Prediksi Kadar Particulate Matter (PM10) untuk Pemantauan Kualitas Udara Menggunakan Jaringan Syaraf Tiruan Studi Kasus Kota Pontianak. *Positron*, *8*(1), 15-20. DOI: http://dx.doi.org/10.26418/positron.v8i1.25470.

[15] Jian, L., Zhao, Y., Zhu, Y. P., Zhang, M. B., & Bertolatti, D. (2012). An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Science of the Total Environment*, *426*, 336-345. DOI: https://doi.org/10.1016/j.scitotenv.2012.03.025.

[16] Miri, A., Shirmohammadi, E., & Sorooshian, A. (2023). Influence of meteorological factors and air pollutants on bacterial concentration across two urban areas of the Sistan region of Iran. *Urban Climate*, *51*, 101650. DOI: https://doi.org/10.1016/j.uclim.2023.101650.

[17] Liu, X., & Guo, H. (2022). Air quality indicators and AQI prediction coupling long-short term memory (LSTM) and sparrow search algorithm (SSA): A case study of Shanghai. *Atmospheric Pollution Research*, *13*(10), 101551. DOI: https://doi.org/10.1016/j.apr.2022.101551.

[18] Wood, D. A. (2022). Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining. *Sustainability Analytics and Modeling*, *2*, 100002.

[19] Zhang, J., & Li, S. (2022). Air quality index forecast in Beijing based on CNN-LSTM multi-model. *Chemosphere*, *308*, 136180. DOI: https://doi.org/10.1016/j.chemosphere.2022.136180.

[20] Udristioiu, M. T., Mghouchi, Y. E., & Yildizhan, H. (2023). Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning. *Journal of Cleaner Production*, *421*, 138496. DOI: https://doi.org/10.1016/j.jclepro.2023.138496.

[21] Luo, J., & Gong, Y. (2023). Air pollutant prediction based on ARIMA-WOA-LSTM model. *Atmospheric Pollution Research*, *14*(6), 101761.

[22] Sengupta, A., Govardhan, G., Debnath, S., Yadav, P., Kulkarni, S. H., Parde, A. N., ... & Ghude, S. D. (2022). Probing into the wintertime meteorology and particulate matter (PM2. 5 and PM10) forecast over Delhi. *Atmospheric Pollution Research*, *13*(6), 101426. DOI: https://doi.org/10.1016/j.apr.2022.101426.

[23] Hardini, M., Sunarjo, R. A., Asfi, M., Chakim, M. H. R., & Sanjaya, Y. P. A. (2023). Predicting air quality index using ensemble machine learning. *ADI Journal on Recent Innovation*, *5*(1Sp), 78-86.

[24] Wang, Y., Huang, L., Huang, C., Hu, J., & Wang, M. (2023). High-resolution modeling for criteria air pollutants and the associated air quality index in a metropolitan city. *Environment International*, *172*, 107752.

[25] Yang, R., & Zhong, C. (2022). Analysis on Spatio-Temporal Evolution and Influencing Factors of Air Quality Index (AQI) in China. *Toxics*, *10*(12), 712.

[26] Jia, N., Li, Y., Chen, R., & Yang, H. (2023). A review of global PM2. 5 exposure research trends from 1992 to 2022. *Sustainability*, *15*(13), 10509.

[27] Li, Y., Chiu, Y. H., & Lu, L. C. (2018). Energy and AQI performance of 31 cities in China. *Energy Policy*, *122*, 194-202. DOI: https://doi.org/10.1016/j.enpol.2018.07.037.

[28] Lunderberg, D. M., Liang, Y., Singer, B. C., Apte, J. S., Nazaroff, W. W., & Goldstein, A. H. (2023). Assessing residential PM2. 5 concentrations and infiltration factors with high spatiotemporal resolution using crowdsourced sensors. *Proceedings of the National Academy of Sciences*, *120*(50), e2308832120. DOI: https://doi.org/10.1073/pnas.2308832120.

[29] Ma, Q., Wang, J., Xiong, M., & Zhu, L. (2023). Air Quality Index (AQI) did not improve during the COVID-19 lockdown in Shanghai, China, in 2022, based on ground and TROPOMI observations. *Remote Sensing*, *15*(5), 1295.

[30] Raaschou-Nielsen, O., Antonsen, S., Agerbo, E., Hvidtfeldt, U. A., Geels, C., Frohn, L. M., ... & Pedersen, C. B. (2023). PM2. 5 air pollution components and mortality in Denmark. *Environment International*, *171*, 107685.

[31] Lin, J., Que, X., Ma, X., Salati, S., Chen, Q., LIU, J., & Fan, C. Spatiotemporal Heterogeneity of the Influence from Main Meteorological Factors on Air Quality Index in China Based on Stwr. *Available at SSRN 4245535*.

[32] Zhao, M., Liu, Y., & Gyilbag, A. (2022). Assessment of meteorological variables and air pollution affecting COVID-19 Cases in urban agglomerations: evidence from China. *International Journal of Environmental Research and Public Health*, *19*(1), 531.

[33] Handhayani, T. (2023). An integrated analysis of air pollution and meteorological conditions in Jakarta. *Scientific Reports*, *13*(1), 5798.

[34] Kartikasari, D. (2020). Analisis Faktor-Faktor yang Mempengaruhi Level Polusi Udara dengan Metode Regresi Logistik Biner. *Mathunesa: Jurnal Ilmiah Matematika*, *8*(1), 55-59. DOI: https://doi.org/10.26740/mathunesa.v8n1.p55-59.

[35] Pratama, I., Permanasari, A. E., Ardiyanto, I., & Indrayani, R. (2016, October). A review of missing values handling methods on time-series data. In *2016 international conference on information technology systems and innovation (ICITSI)* (pp. 1-6). IEEE. DOI: 10.1109/ICITSI.2016.7858189.

[36] Ahn, H., Sun, K., & Kim, K. P. (2022). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials & Continua*, *70*(1), 767-779.

[37] Adu, W. K., Appiahene, P., & Afrifa, S. (2023). VAR, ARIMAX and ARIMA models for nowcasting unemployment rate in Ghana using Google trends. *Journal of Electrical Systems and Information Technology*, *10*(1), 12.

[38] Ospina, R., Gondim, J. A., Leiva, V., & Castro, C. (2023). An overview of forecast analysis with ARIMA models during the COVID-19 pandemic: Methodology and case study in Brazil. *Mathematics*, *11*(14), 3069.

[39] Islam, F., & Imteaz, M. A. (2020). Use of teleconnections to predict Western Australian seasonal rainfall using ARIMAX model. *Hydrology*, *7*(3), 52.

[40] Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*, *10*, 1847979018808673. DOI: https://doi.org/10.1177/1847979018808673.

[41] Alabdulrazzaq, H., Alenezi, M. N., Rawajfih, Y., Alghannam, B. A., Al-Hassan, A. A., & Al-Anzi, F. S. (2021). On the accuracy of ARIMA based prediction of COVID-19 spread. *Results in Physics*, *27*, 104509.

[42] Klabi, F. (2014). Forecasting non-residents' monthly entries to Tunisia and accuracy comparison of time-series methods. *The Journal of North African Studies*, *19*(5), 770-791. DOI: https://doi.org/10.1080/13629387.2014.949694.