BAREKENG: Journal of Mathematics and Its Applications
December 2026 Volume 20 Issue 4 Page 2825-2838
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss4pp2825-2838

INTEGRATION OF QUADRATIC REGRESSION-ARIMA MODEL ESTIMATING AIR QUALITY INDEX ON PM2.5 CONCENTRATION

Tiara Herlinda Sari [1*], Yundari [2], Shantika Martha [3]

[1,2,3] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura*
Jln. Prof. Dr. H. Hadari Nawawi, Pontianak, 78124, Indonesia

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

Article Info

Article History:
Received: 10th September 2025
Revised: 20th February 2026
Accepted: 22nd May 2026
Available online: 1st July 2026

Keywords:
air pollution;
AQI;
ARIMA;
estimation;
regression.

ABSTRACT

In Air Quality Index (AQI) computation, the relationship between AQI and PM2.5 is defined through an interpolation approach, which forms a nonlinear relationship between PM2.5 and AQI, thereby rendering linear regression less capable of representing this relationship. This research selects quadratic regression because it explicitly represents the nonlinear relationship between PM2.5 and AQI while remaining easy to interpret and minimizing the risk of overfitting compared with more complex nonlinear models. However, in time series data, this model still generates residuals that violate classical assumptions due to time dependence. Hence, a hybrid modeling approach integrates quadratic regression and Autoregressive Integrated Moving Average (ARIMA) to represent nonlinear relationships while correcting the time dependence on the quadratic regression residuals. This research aims to estimate AQI based on PM2.5 in Pontianak in 2024 using the Quadratic Regression-ARIMA model, also comparing the performance of quadratic regression models and quadratic regression-ARIMA models. This research focuses on estimation and interpolation within the sample, not forecasting. The dataset consists of 366 daily observations of AQI and PM2.5 throughout 2024, obtained from the official air quality monitoring website, AQI. The analysis was carried out by estimating the AQI from PM2.5 using a Quadratic Regression model, and the regression residuals were rendered stationary before ARIMA modeling. The results showed that the Quadratic Regression-ARIMA model yields estimate with a better residual structure than the quadratic regression model. The in-sample evaluation of the Quadratic Regression-ARIMA model showed better performance, with a higher coefficient of determination (R^2) of 0.95, compared with 0.91 for quadratic regression and 0.80 for linear regression.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

T. H. Sari, Yundari and S. Martha., "INTEGRATION OF QUADRATIC REGRESSION-ARIMA MODEL ESTIMATING AIR QUALITY INDEX ON PM2.5 CONCENTRATION", *BAREKENG: J. Math. & App.,* vol. 20, no. 4, pp. 2825-2838, Dec, 2026.

1. INTRODUCTION

Air pollution has emerged as a major environmental concern, exerting a substantial influence on overall air quality [1]. This is evidenced by the fact that over 90% of the global population inhabits regions where air quality levels fall below the thresholds recommended by the World Health Organization (WHO), so the assessment of air quality necessitates the use of appropriate indicators [2]. One established metric is the Air Quality Index (AQI). A higher index value indicates higher pollutant concentrations, so air quality in the area is worsening, whereas lower index values indicate better air quality [3]. Consequently, accurate AQI estimates are essential for supporting environmental policies, public health early warning systems, and urban air quality assessments [4].

The concentrations of various air contaminants are used to determine the AQI, including PM2.5, SO_2, NO_2, CO, and O_3 [3]. Among these pollutants, PM2.5 is a major determinant of the AQI and poses the most serious threat to human health [5]. Its extremely small size, with a diameter of 2.5 µm or less, enabling it to reach the deepest parts of the lungs and pass into the circulatory system [6]. In Pontianak, PM2.5 has become an environmental problem due to the tropical climate, urban activities, and frequent forest and land fires [7] which produce haze, thereby increasing PM2.5 concentrations and causing fluctuations in the AQI [8]. The relationship between PM2.5 and AQI is defined by interpolation between the upper and lower limits of PM2.5 and AQI. This approach demonstrates a nonlinear relationship between PM2.5 and AQI due to the division of index intervals, and PM2.5 intervals vary across concentration ranges [9]. Hence, the linear regression approach is less capable of representing this relationship.

Previous studies have used various nonlinear approach to air quality modeling, such as kernel regression [10], support vector regression (SVR) [11], and machine learning methods such as Random Forest [12]. However, these methods are typically prediction-oriented and involve more complex model structures, meaning that the functional relationships between variables are not explicitly defined mathematically. This research focuses on developing simple non-linear models that allow for clear mathematical interpretation. Hence, quadratic regression was selected because it explicitly represents the nonlinear relationship between PM2.5 and AQI while remaining easy to interpret and minimizing the risk of overfitting compared with more complex nonlinear models [13]. Meanwhile, the ARIMA model was preferred to other methods because it effectively captures and corrects temporal dependence in the residuals through the autoregressive and moving average components. In addition, ARIMA provides well-established diagnostic tests to verify that autocorrelation structures have been adequately addressed. Accordingly, integrating the Quadratic Regression and ARIMA models is expected to enhance the performance of AQI estimates [14].

The limitations of the modeling approach in this research are the use of daily observations throughout 2024 with a total of 366 observations, the restriction to a single location, and the consideration of only one type of pollutant, namely PM2.5. However, this approach remains relevant due to the dominant role of PM2.5 in urban air pollution, including in Pontianak, and its significant impact on public health [8]. Therefore, this research aims to estimate AQI based on PM2.5 in Pontianak in 2024 using the Quadratic Regression-ARIMA model, and to compare the performance of quadratic regression models and quadratic regression-ARIMA models. A hybrid modeling approach that combines regression and time series models has also been used in various studies [14]. The novelty of this research is the integration of Quadratic Regression-ARIMA for PM2.5-based AQI estimation, capturing the nonlinear relationship between PM2.5 and AQI while modeling the remaining temporal dynamics in the data. Therefore, this research is expected to make new contributions to the development of hybrid regression and time series models.

2. RESEARCH METHODS

The analysis process for this research began with descriptive statistical analysis and regression modeling. Since the relationship between the AQI and PM2.5 is by definition nonlinear, quadratic regression was used. However, linear regression was also evaluated as a comparison for model performance. The estimation results from the quadratic regression model produce residuals that violate the specified assumptions, which motivates further analysis by treating the residuals as a time series and modeling them using an ARIMA model. Thus, a quadratic regression-ARIMA model was used. After identifying the optimal ARIMA specification through standard ARIMA procedures, this research estimates AQI by integrating the quadratic regression and ARIMA models, combining their estimated components, and reassessing the resulting residuals to evaluate whether the integration enhances model performance.

2.1 Quadratic Regression Model

quadratic regression model is a nonlinear regression technique in which the relationship between the independent and dependent variables is modeled by a parabolic curve [15]. However, this model remains linear in its parameters and can be expressed as a linear regression model by transforming the variables [16]. The formulation of the model is presented below[17]:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t, \tag{1}$$

with:

Y_t : dependent variable for t;

X_t : independent variable for t;

β_0 : constant;

β_1 : linear coefficient;

β_2 : quadratic coefficient;

ε_t : residual for t.

2.2 Estimation of Quadratic Regression Model Parameters

Estimation of quadratic regression model parameters using the least squares method, which reduces the sum of squared residuals of regression model equation. The general equation for estimating parameters β_0, β_1, and β_2 can be obtained as follows [18]:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} n & \sum_{t=1}^{n} X_t & \sum_{t=1}^{n} X_t^2 \\ \sum_{t=1}^{n} X_t & \sum_{t=1}^{n} X_t^2 & \sum_{t=1}^{n} X_t^3 \\ \sum_{t=1}^{n} X_t^2 & \sum_{t=1}^{n} X_t^3 & \sum_{t=1}^{n} X_t^4 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{n} Y_t \\ \sum_{t=1}^{n} Y_t X_t \\ \sum_{t=1}^{n} Y_t X_t^2 \end{bmatrix}, \tag{2}$$

with:

$\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$: parameter estimates;

n : the amount of data.

2.3 Model Fit Test

Model fit can be assessed using the F- and t-tests, along with the coefficient of determination (R^2) [19]. The combined impact of the independent variables on the dependent variable is assessed using the F-test for significance. In contrast, the t-test focuses on evaluating the effect of each variable individually. The coefficient of determination (R^2) has a value between 0 and 1; the closer the value is to 1, the better the model explains the variation in the dependent variable. It is calculated by dividing between the Sum of Squares for Regression (SSR) and Total Sum of Squares (SST) [20]. The F significance test requires the Mean Square Regression (MSR) and Mean Square Error (MSE) values using equations adopted from [21]:

$$MSR = \frac{\sum_{t=1}^{n} \left(\hat{Y}_t - \bar{Y}\right)^2}{k}, \tag{3}$$

$$MSE = \frac{\sum_{t=1}^{n} (Y_t - \bar{Y})^2 - \sum_{t=1}^{n} \left(\hat{Y}_t - \bar{Y}\right)^2}{n - k - 1}. \tag{4}$$

The following F-test, t-test, and coefficient of determination are calculated using equations adopted from [21]:

$$F_{calc} = \frac{MSR}{MSE}, \tag{5}$$

$$t_{calc} = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}, \tag{6}$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}, \tag{7}$$

with:

$\hat{\beta}_i$: estimate for β_i;

$SE(\hat{\beta}_i)$: standard error for $\hat{\beta}_i$;

Y_t : dependent variable for t or actual data for t;

$\hat{Y}_t$: estimated value for t;

$\bar{Y}$: average of the actual data.

F-test to examine the null hypothesis that the independent variables collectively do not exert a statistically significant influence. In the test criterion, when the resulting p-value exceeds the significance level of 5%, the test rejects the null hypothesis, indicating that the independent variables collectively do not exert a statistically significant influence. Conversely, a p-value below the significance level leads to rejection of the null hypothesis, implying that at least one independent variable collectively exerts a statistically significant influence. In addition, a t-test is used to test the null hypothesis that a specific independent variable has no statistically significant effect on the dependent variable. In the test criterion, when the resulting p-value exceeds the significance level of 5%, the test accepts the null hypothesis, indicating the absence of a significant individual effect; a p-value below the significance level results in rejection of the null hypothesis, indicating the presence of a statistically significant individual effect.

2.4 Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA (p, d, q) formulation, the model involves three parameters, namely AR(p) which denotes autoregressive part, MA(q) which denotes the moving average part, and d which denotes the level of difference [22]. ARIMA specification with parameters p, d, dan q on Y_t data are presented in the following equation as adopted from [23]:

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)a_t, \tag{8}$$

with:

$(1-B)^d$: the order differentiation for d;

$\phi_p(B)$: the AR coefficient, with $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$;

$\theta_q(B)$: the MA coefficient, with $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$;

a_t : the residual for t.

Before determining ARIMA (p, d, q), the procedure begins by assessing the data for stationarity, namely, the stationarity of both the mean and the variance. Fulfilling the stationarity assumption is important to ensure parameter stability and estimation accuracy in ARIMA modelling [24]. If not stationary in variance with a lambda (λ) value not equal to 1, then a Box-Cox transformation will be performed as follows [24]:

$$T(Y_t) = \begin{cases} \dfrac{Y_t^{(\lambda)} - 1}{\lambda}, & \lambda \neq 0, \\ \ln Y_t, & \lambda = 0 \end{cases} \tag{9}$$

The lambda (λ) value is used to stabilize the data variance. In cases where the data does not exhibit mean stationarity, then a differentiation process will be performed, which is expressed by the backshift operator (B) as follows [25]:

$$B^d Y_t = Y_{t-d} \quad , d > 0. \tag{10}$$

The Augmented Dickey-Fuller (ADF) test assesses stationarity at the mean, as presented in the following equation [24]:

$$ADF_{calc} = \frac{\hat{\rho}}{SE(\hat{\rho})}, \quad (11)$$

with $\hat{\rho}$ denotes the estimated autoregressive coefficient parameter that indicates the presence of unit roots in the time series, whereas $SE(\hat{\rho})$ represents the corresponding standard error of the estimate. In the ADF test, the null hypothesis is that the series is non-stationary in the mean at the 5% significance level; if the p-value is less than 0.05, the null hypothesis is rejected, indicating mean stationarity, whereas a p-value greater than the significance level denotes the opposite. Next, assessment of autocorrelation and partial autocorrelation plots for model identification within the ARIMA framework, then parameter estimation and significance testing. The procedure tests the null hypothesis that each estimated parameter does not contribute significantly to the model. The null hypothesis is rejected if the p-value is below 5%, which means that the parameter contributes significantly to the ARIMA model [24].

Subsequently, perform diagnostic tests to ensure that the ARIMA residuals meet the basic assumptions. The Ljung-Box test tests the null hypothesis that the residuals are white noise at the 5% significance level; if the p-value is greater than 0.05, the null hypothesis is accepted, indicating that the ARIMA model residuals are white noise [26]. Meanwhile, the ARCH-Lagrange Multiplier (ARCH-LM) test examines the null hypothesis that the variance of the residuals is constant (i.e., homoscedasticity) or that the residuals exhibit no ARCH effect. At a significance level of 5%, if p-values are greater than 0.05, the null hypothesis of homoscedasticity is accepted [27]. In addition to statistical tests, residuals were also evaluated visually using residual autocorrelation function (ACF) plots and Q-Q plots to verify independence and normality [28]. The most appropriate ARIMA model can be identified by evaluating model performance using the Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), and Mean Absolute Error (MAE). The optimal ARIMA specification is achieved when its performance indicators reach the smallest values [25].

2.5 Quadratic Regression-ARIMA Model

The quadratic regression-ARIMA model starts by applying a quadratic regression model to the data, then modeling the residuals from the quadratic regression that still contain time patterns with an ARIMA model, and finally combining the estimates from the quadratic regression and the ARIMA model. The Quadratic Regression-ARIMA equation is presented below:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \eta_t, \quad (12)$$

with η_t is residual component of the Quadratic Regression model modeled using the ARIMA (p, d, q) as follows [29]:

$$\phi_p(B)(1 - B)^d \eta_t = \theta_q(B) a_t. \quad (13)$$

The analytical procedure in this research began with descriptive statistics, followed by quadratic regression to estimate parameters. Subsequent stages included evaluating model adequacy using the F and t significance tests, along with the coefficient of determination, and estimating AQI. Further analysis was conducted by applying ARIMA modeling to the residuals. This procedure involved testing for stationarity, identifying an ARIMA model using ACF and PACF (Partial Autocorrelation Function) plots, estimating model parameters, selecting the most suitable ARIMA configuration, and performing diagnostic checks on the residuals. After determining the optimal ARIMA model, the quadratic regression residuals are estimated using it. Next, in-sample estimation is obtained by integrating the quadratic regression-ARIMA, in which quadratic regression residual estimates are processed using ARIMA added to the AQI estimates from the quadratic regression model. The residuals from the model were then re-evaluated to assess the integrated model's reliability and estimation performance [29].

2.6 Model Performance Evaluation

The performance of the models within the sample was evaluated using the coefficient of determination (R^2). This method is selected because the analysis emphasizes the models' capability to estimate and interpolate over the observation period. The R^2 value serves to compare the performance of the quadratic regression model and the Quadratic Regression-ARIMA model with R^2 values closer to 1 represent better model performance [20].

3. RESULTS AND DISCUSSION

The AQI is the dependent variable (Y), while the PM2.5 is stated as the independent variable (X). An illustration of the connection between the two variables is provided in Fig. 1.

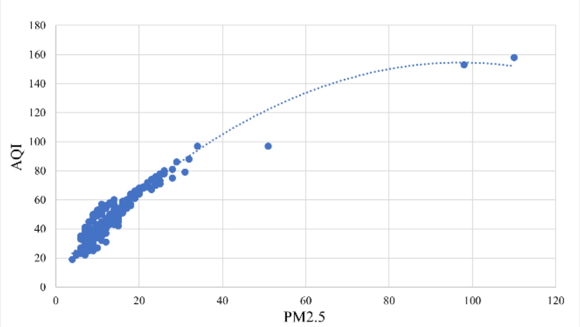


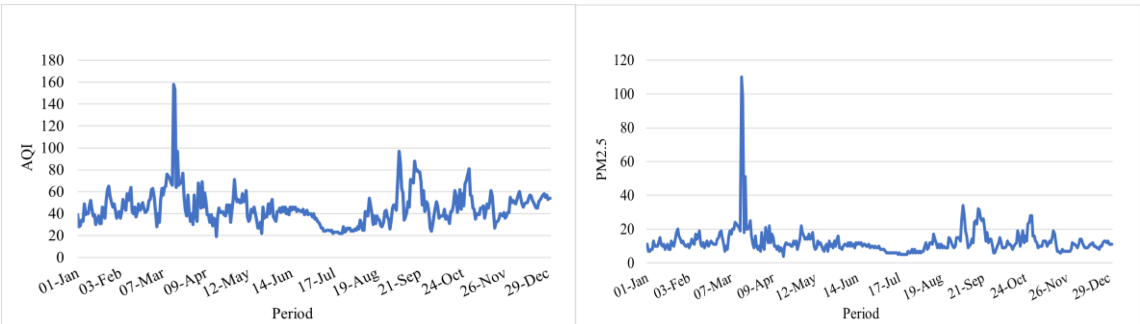
Figure 1. Scatter Plot of Air Quality Index and PM2.5

The distribution of data points in Fig. 1, indicates that increasing PM2.5 is associated with increasing AQI, so the data distribution appears to form a curved pattern, indicating a nonlinear relationship among variables. An overview of the data characteristics is provided in Table 1.

Table 1. Descriptive Statistics

Variable	n	Minimum	Maximum	Average	Standard Deviation
Y	366	19	158	45.178	15.781
X	366	4	110	12.593	8.615

The lowest AQI value falls within the good AQI category, which occurred on April 17, while the highest AQI value falls within the unhealthy AQI category, which occurred on March 15. The lowest and highest PM2.5 concentrations also occurred on the same days as the AQI, with the lowest falling within the good category and the highest within the unhealthy category. The average AQI value for the entire year of 2024 is 45.178 with a standard deviation of 15.781, while the average PM2.5 is 12.593 $\mu g/m^3$ with a standard deviation of 8.615. The time series data plot is shown in Fig. 2.



(a) (b)

Figure 2. (a) Plot Data for Air Quality Index and (b) Plot Data for PM2.5 Concentration

As shown in Fig. 2, the AQI and PM2.5 plots exhibit daily fluctuations throughout 2024, with AQI increasing significantly in early March, then declining mid-year before rising again toward the end of the year. Fluctuations in PM2.5 follow the same pattern as AQI. This may be attributed to external factors such as combustion activities or weather conditions.

3.1 Formation of Quadratic Regression Model

The initial step in forming a quadratic regression model is to estimate the parameters using the formula in Eq. (2), which yields $\hat{\beta}_0 = 11.705$; $\hat{\beta}_1 = 2.937$; and $\hat{\beta}_2 = -0.015$, so that a quadratic regression estimation model can be formed as follows:

$$\hat{Y}_t = 11.705 + 2.937X_t - 0.015X_t^2. \tag{14}$$

Next, model fit testing will be conducted using F and t significance tests, along with along with the coefficient of determination. The outcome of the F-test calculations using the formula in Eq. (5) is tabulated in Table 2.

Table 2. F-test

***F* statistic**	**p-value**
1768.904	0.000

The independent factors together exert a significant influence on the dependent variable; in Table 2, at the 5% significance level, this yields an F-statistic of 1768.904 and a p-value less than 0.05, leading to the rejection of the null hypothesis. Next, a t-test was performed using the formula in Eq. (6), and the results are presented in Table 3.

Table 3. T-test

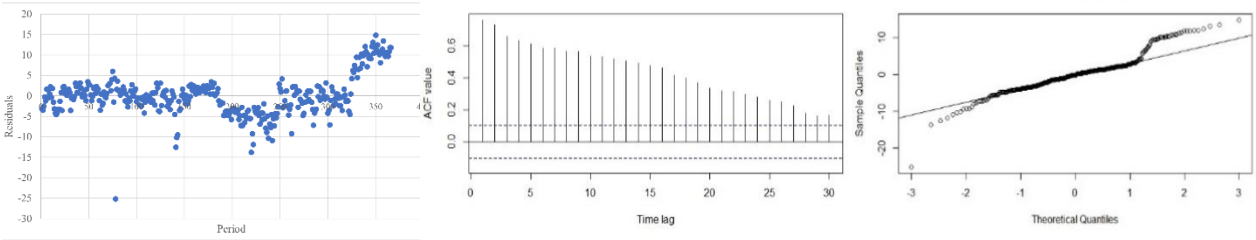
Variable	***t* statistic**	**p-value**
X	42.311	0.000
X^2	-20.687	0.000

As presented in Table 3, at the 5% significance level, the variables X and X^2 have p-values of 0.000, leading to the rejection of the null hypothesis. These results confirm that each independent variable individually exerts a statistically significant influence on the dependent variable. Next, the coefficient of determination calculated using Eq. (7) is presented in Table 4.

Table 4. Coefficient of Determination

R^2	**Adjusted R^2**
0.91	0.91

As presented in Table 4, the coefficient of determination (R^2) is 0.91, indicating that the model explains 91% of the variation in the dependent variable, with the remaining 9% attributable to external factors. Meanwhile, the adjusted coefficient of determination is also 0.91, indicating that the model still explains 91% of the variation in the dependent variable, given the number of variables and the sample size. The residual test is shown in Fig. 3.



(a) (b) (c)

Figure 3. (a) Residual Pattern of Quadratic Regression Model, (b) ACF Residuals Plot of Quadratic Regression Model, and (c) Q-Q Plot Residuals of Quadratic Regression Model

As presented in Fig. 3, the residual pattern in Fig. 3 (a), formed from residual points that spread out in a non-random pattern and deviate from the zero line, suggests the presence of unexplained information not fully captured within the model. Therefore, the assumptions regarding the randomness of residuals are not met. The residual autocorrelation function plot in Fig. 3 (b) shows that the autocorrelation values at each lag (lag 1, lag 2, etc) are outside the significance line (dashed line). This indicates that the model has not fully captured the time pattern in the data, and the residuals still contain a temporal pattern. Based on the Q-Q plot in Fig. 3 (c), the residuals deviate from the expected diagonal, indicating a violation of assumptions due to several outliers in the dataset, so the residuals are not normally distributed. Based on the residual test, the residual components of the quadratic regression model are then modeled using ARIMA.

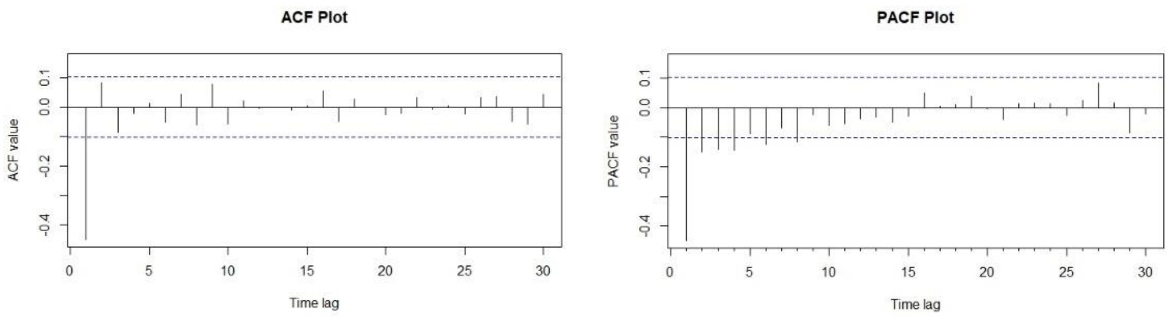
3.2 Formation of ARIMA Model

Residual outcomes from the quadratic regression are subsequently analyzed with an ARIMA model to account for temporal dependencies. The initial stage in constructing the ARIMA model involves examining whether the data exhibit stationarity. Testing for variance stability yielded a lambda (λ) value of 1.03, indicating that the data exhibit variance stationarity, so the procedure can proceed to evaluate mean stationarity through the Augmented Dickey-Fuller (ADF) approach in Eq. (11) refer to Table 5.

Table 5. ADF Test

Statistic value	p-value	Stationarity status
Before differencing	0.65	Non-stationary
After differencing	0.01	Stationary

As shown in Table 5, at the 5% significance level, the p-value before the differentiation process exceeds 0.05, leading to acceptance of the null hypothesis and indicating that the data lack stationarity in the mean. Therefore, a differentiation process was necessary. After the differentiation process, a p-value below 0.05 was obtained, leading to rejection of the null hypothesis and indicating the presence of mean stationarity in the data. Identification of the ARIMA structure was carried out by examining ACF and PACF plots of the residuals that satisfied stationarity, as shown in Fig. 4.



(a) (b)

Figure 4. (a) ACF Plot and (b) PACF Plot

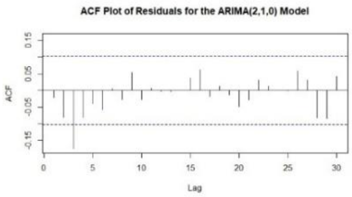
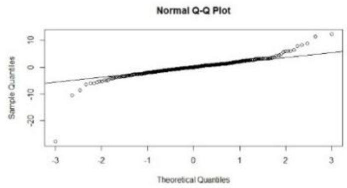
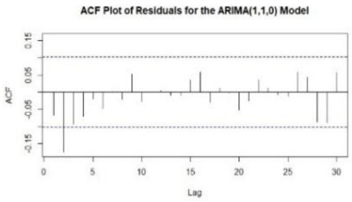
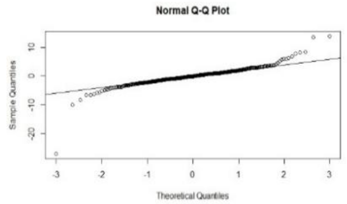
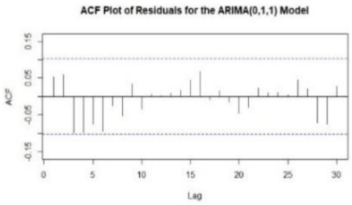
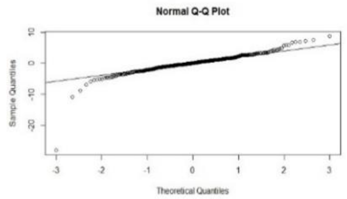
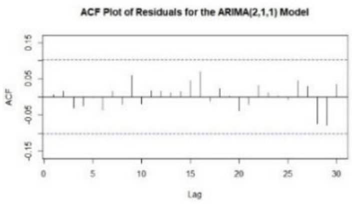
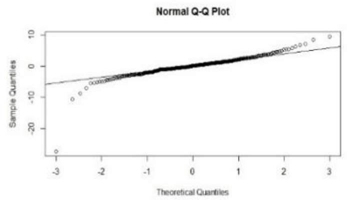
As presented in Fig. 4, truncation after the first lag is evident on the ACF plot, whereas the PACF plot displays evident truncations at the first and second lags. For parameter d, its value is determined as 1 because the data were differentiated once. Based on both plots, four temporary ARIMA models are obtained: ARIMA (2,1,0), ARIMA (1,1,0), ARIMA (0,1,1), and ARIMA (2,1,1). Considering parameters significant, model estimation is conducted if they have a p-value below 0.05, thereby qualifying them for integration into the ARIMA model. A detailed summary of the estimated ARIMA parameters is presented in Table 6.

Table 6. Estimation of ARIMA Model Parameters

Model	Parameter		p-value
ARIMA (2,1,0)	ϕ_1	-0.514	0.000
	ϕ_2	-0.147	0.004
ARIMA (1,1,0)	ϕ_1	-0.448	0.000
ARIMA (0,1,1)	θ_1	-0.640	0.000
ARIMA (2,1,1)	ϕ_1	0.252	0.000
	ϕ_2	0.171	0.005
	θ_1	-0.865	0.000

As shown in Table 6, at the 5% significance level, each ARIMA model parameter has a p-value below 0.05, leading to rejection of the null hypothesis. This means that all parameters contribute significantly to the ARIMA model. Subsequently, each significant ARIMA model is evaluated through diagnostic tests, as summarized in Table 7.

Table 7. Diagnostic Test

Model	Ljung-Box test (p-value)	ARCH-LM test (p-value)	Residual independence	Residual normality
ARIMA (2,1,0)	0.286	0.876		
ARIMA (1,1,0)	0.236	0.124		
ARIMA (0,1,1)	0.575	1		
ARIMA (2,1,1)	0.985	0.999		

As shown in Table 7 the Ljung-Box test at the 5% significance level yields p-values greater than 0.05 for all ARIMA models, leading to acceptance of the null hypothesis. This result indicates that the residuals from all ARIMA models do not exhibit simultaneous autocorrelation, and the white noise assumption is satisfied. Visually, the ACF residual plots show that in the ARIMA (2,1,0) model, the autocorrelation value at lag 3 exceeds the significance line (dashed line), while in the ARIMA (1,1,0) model, the autocorrelation value at lag 2 also exceeds the significance line (dashed line). This discrepancy arises because the ACF plot provides a lag-by-lag visual assessment of autocorrelation. In contrast, the Ljung-Box test assesses simultaneous autocorrelation over a specified number of lags, so the ARIMA model still satisfies the white noise assumption. Conversely, the ARIMA (0,1,1) and ARIMA (2,1,1) models satisfy the white noise assumption both visually and statistically, as their autocorrelation values at all lags are within the significance level.

Furthermore, in the ARCH-LM test at a 5% significance level, all p-values are greater than 0.05, so the null hypothesis is accepted, representing that the residuals for all ARIMA models satisfy the homoscedasticity assumption. Additionally, the residual normality plots show that the points generally align closely with the diagonal, suggesting a normal distribution. Hence, all ARIMA models satisfy the diagnostic test. Furthermore, the optimal ARIMA model is determined by evaluating RMSE, AIC, and MAE, where smaller values indicate a better model. The AIC, RMSE, and MAE values are summarized in Table 8.

Table 8. Best Model Selection

Model	AIC	RMSE	MAE
ARIMA (2,1,0)	1817	2.88	1.86
ARIMA (1,1,0)	1823.04	2.91	1.88
ARIMA (0,1,1)	1798.29	2.82	1.86
ARIMA (2,1,1)	1789.53	2.77	1.82

As presented in Table 8, the smallest AIC, RMSE, and MAE values were achieved by the ARIMA (2,1,1) model, identified as the optimal model, whose equation is formulated as follows:

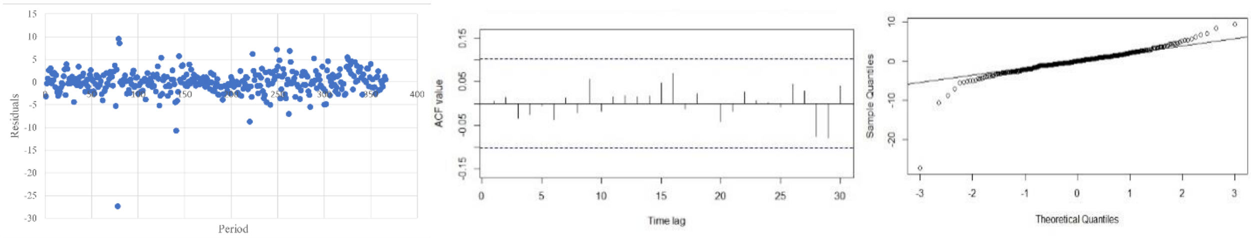
$$\eta_t = 1.252\eta_{t-1} - 0.081\eta_{t-2} - 0.171\eta_{t-3} + a_t + 0.865a_{t-1} \tag{15}$$

3.3 Quadratic Regression-ARIMA Model

After obtaining the Quadratic Regression model and ARIMA model from Quadratic Regression residuals, the next step is to form a Quadratic Regression-ARIMA model by combining the two models. Based on Eqs. (14) and (15), the Quadratic Regression-ARIMA estimation model is as follows:

$$\hat{Y}_t = 11.705 + 2.937X_t - 0.015X_t^2 + 1.252\eta_{t-1} - 0.081\eta_{t-2} - 0.171\eta_{t-3} + 0.865a_{t-1}.$$

This model shows that when PM2.5 equals zero, the estimated AQI in period t is 11.705, indicating that the estimated AQI is unaffected by PM2.5 but remains influenced by other pollutants. Accordingly, even though the PM2.5 value is 0, the estimated result still has a value because the estimate reflects the influence of other pollutants or environmental factors that the model does not explicitly incorporate. When the PM2.5 increases by 1 $\mu g/m^3$, the estimated AQI value is expected to increase by 2.937. Still, this effect does not continue because there is a negative coefficient of 0.015 on the square of PM2. 5, which indicates that after exceeding a concentration of approximately 98 $\mu g/m^3$, an increase in PM2.5 has a negative impact on the estimated AQI. At the same time, the previous day's influence also affects the estimated AQI, such that if the estimated AQI was higher the previous day, today's estimate tends to increase by 1.252, but if the model overestimated the previous two or three days, today's estimate will decrease by 0.081 and 0.171, respectively. If the model underestimated the previous day, it will adjust today's estimate upward by 0.865 to correct for the error. The next step is to examine the residuals to evaluate characteristics of the residuals produced by the model. The residual test can be observed in Fig. 5.



(a) (b) (c)

Figure 5. Residual Pattern of Quadratic Regression-ARIMA Model, (b) ACF Residuals Plot of Quadratic Regression-ARIMA Model, and (c) Q-Q Plot Residuals of Quadratic Regression-ARIMA Model

As presented in Fig. 5, the residual pattern in Fig. 5 (a) consists of residual points dispersed randomly around zero without creating any particular pattern, suggesting the absence of association between the residuals and either time or the predictor variables. Therefore, the assumption of residual randomness is satisfied, and the model is good at estimating. The residual autocorrelation function plot in Fig. 5 (b) demonstrates that the autocorrelation values at each lag (lag 1, lag 2, etc) are within the significance line (dashed line), which indicates that there is no significant autocorrelation between the residuals. Thus, it satisfies the assumption of residual independence, in which the residuals do not contain time. Based on the Q-Q plot in Fig. 5 (c), the data distribution approaches the diagonal line, indicating that the residual component conforms to a normal distribution. Therefore, the assumption of residual normality is met. These results demonstrate that the quadratic regression-ARIMA model yields better residual characteristics than those before the addition of the ARIMA component. Thus, integrated quadratic regression and ARIMA demonstrate effective capability in identifying patterns within the data. However, since the approach is carried out gradually, the final results still contain cumulative errors originating from both model components.

3.4 Model Performace Evaluation

The performance evaluation of the quadratic regression model and the quadratic regression-ARIMA model was performed using the R^2 value. As an additional comparison, the results from the linear regression were also presented, which clarified the difference in performance between the linear and nonlinear models, as presented in Table 9.

Table 9. Model Performance Evaluation

Model	R^2
Linear Regression	0.80
Quadratic Regression	0.91
Quadratic Regression-ARIMA	0.95

As shown in Table 9, the linear regression model yielded an R^2 of 0.80. In contrast, the quadratic regression model yielded an R^2 value of 0.91, indicating that the nonlinear approach better explains the variation in the air quality index than linear regression. Furthermore, the quadratic regression-ARIMA model exhibited the highest R^2 value of 0.95. This increase in the R^2 value indicates that the addition of the ARIMA component captured the time dependence of the residuals in the quadratic regression model, thereby improving estimation performance. Findings from this study reveal that PM2.5 exerts a notable influence on the AQI. This is consistent with previous research findings indicating that PM2.5 is one of the factors significantly influencing changes in the AQI [6].

4. CONCLUSION

From the analytical outcomes, it may be inferred that the Quadratic Regression-ARIMA model is a relevant approach for estimating the AQI based on PM2.5 in Pontianak City for 2024. However, since this approach focuses on interpolation and estimation within the sample, the results obtained are limited to the data and time period of this study. The use of this hybrid model is motivated by the nonlinear relationship and temporal dependence inherent in the data. The results showed that the Quadratic Regression-ARIMA model yields estimate with a better residual structure than the quadratic regression model. The in-sample evaluation of the Quadratic Regression-ARIMA model showed better performance, with a higher R^2 of 0.95, ompared with 0.91 for quadratic regression and 0.80 for linear regression. This improvement in the hybrid model's performance occurred because the ARIMA component was able to model the time dependence in the residuals that could not be fully explained by quadratic regression. This approach can support environmental policy-making and air quality monitoring. However, this research has several limitations, including the use of daily observations throughout 2024, totaling 366 observations, the restriction to a single location, and the consideration of only one type of pollutant. Further research is recommended to explore other modeling methods, such as nonparametric approaches, ensemble models, Long Short-Term Memory (LSTM), and to include additional pollutants and different locations to enhance the generalizability and performance of models in future studies.

Author Contributions

Tiara Herlinda Sari: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing - Original Draft, Writing - Review and Editing. Yundari: Conceptualization, Funding Acquisition, Methodology, Supervision, Validation, Writing - Review and Editing. Shantika Martha: Conceptualization, Funding Acquisition, Methodology, Supervision, Validation, Writing - Review and Editing. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This investigation was carried out without funding from governmental bodies, private enterprises, or nonprofit organizations.

Acknowledgment

The authors would like to convey their sincere appreciation and gratitude to everyone who has provided support the completion of this work.

Declarations

The authors declare no conflicts of interest to report study.

Declaration of Generative AI and AI-assisted technologies

AI-assisted technology (ChatGPT) was used to support sentence restructuring and clarity improvements. The authors confirm that the underlying ideas, arguments, data analyses, and conclusions are original and were not generated by AI. All AI-assisted edits were critically reviewed and validated by the authors.

REFERENCES

[1] V. Rahmawati, A. L. Hayat, and A. Salam, "ANALISIS DAMPAK PENCEMARAN UDARA TERHADAP KESEHATAN MASYARAKAT DI PERKOTAAN," *SEMAR*, vol. 2, no. 3, pp. 17–24, 2024, doi: https://doi.org/10.59966/semar.v2i3.885.

[2] X. Li and X. Liu, "EFFECTS OF PM2.5 ON CHRONIC AIRWAY DISEASES: A REVIEW OF RESEARCH PROGRESS," *Atmosphere*, vol. 12, no. 8, p. 1068, 2021, doi: https://doi.org/10.3390/atmos12081068.

[3] R. Yang and C. Zhong, "ANALYSIS ON SPATIO-TEMPORAL EVOLUTION AND INFLUENCING FACTORS OF AIR QUALITY INDEX (AQI) IN CHINA," *Toxics*, vol. 10, no. 12, p. 712, 2022, doi: https://doi.org/10.3390/toxics10120712.

[4] T. Liu and S. You, "ANALYSIS AND FORECAST OF BEIJING'S AIR QUALITY INDEX BASED ON ARIMA MODEL AND NEURAL NETWORK MODEL," *Atmosphere*, vol. 13, no. 4, p. 512, 2022, doi: https://doi.org/10.3390/atmos13040512.

[5] I. Wellid, S. Sugiyarto, L. M. Simbolon, G. A. Nur Fajrina, N. Nuryati, and M. F. Bin Sukri, "AIR POLLUTANT OF PM10 AND PM2.5 ON AIR QUALITY INDEX IN BUSINESS CENTER KOTA BARU PARAHYANGAN, WEST BANDUNG DISTRICT," *IJATR*, vol. 5, no. 2, pp. 97–106, 2024, doi: https://doi.org/10.35313/ijatr.v5i2.163.

[6] D. K. Nurlaili and N. Hendrasarie, "ANALISA KUALITAS LINGKUNGAN UDARA AMBIEN (PM2.5) DI KOTA SURABAYA," *JSE*, vol. 9, no. 1, pp. 7988–7995, 2023, doi: https://doi.org/10.32672/jse.v9i1.713.

[7] M. F. Bafadal and A. P. Hestiantini, "Kebijakan Pemerintah Kalimantan Barat Dalam Isu Lingkungan Kebakaran Hutan Dan Lahan (Karhutla) Tahun 2023," *INNOVATIVE: Journal Of Social Science Research*, vol. 3, no. 6, pp. 4528–4538, 2023.

[8] K. H. Cheong *et al.*, "ACUTE HEALTH IMPACTS OF THE SOUTHEAST ASIAN TRANSBOUNDARY HAZE PROBLEM—A REVIEW," *IJERPH*, vol. 16, no. 18, p. 3286, 2019, doi: https://doi.org/10.3390/ijerph16183286.

[9] R. Rita, D. Dwiana Lestiani, E. Hamonangan Panjaitan, M. Santoso, and H. Yulinawati, "KUALITAS UDARA (PM10 DAN PM2.5) UNTUK MELENGKAPI KAJIAN INDEKS KUALITAS LINGKUNGAN HIDUP," *ecolab*, vol. 10, no. 1, pp. 1–7, 2016, doi: https://doi.org/10.20886/jklh.2016.10.1.1-7

[10] A. Donnelly, B. Misstear, and B. Broderick, "REAL TIME AIR QUALITY FORECASTING USING INTEGRATED PARAMETRIC REGRESSION TECHNIQUES," *Elsevier*, vol. 103, pp. 53–65, 2015.doi: https://doi.org/10.1016/j.atmosenv.2014.12.011

[11] R. Wahyudi, S. Annas, and Z. Rais, "ANALISIS SUPPORT VECTOR REGRESSION (SVR) UNTUK MERAMALKAN INDEKS KUALITAS UDARA DI KOTA MAKASSAR," vol. 5, pp. 104–117, 2023.

[12] A. Rivaldi, A. Ramadhani, I. Ramadhan, and Y. Nuryamin, "PREDIKSI DAN PEMODELAN KUALITAS UDARA MENGGUNAKAN RANDOM FOREST DAN GRADIENT BOOSTING JAKARTA DAN TANGERANG," vol. 5, no. 4, 2025.

[13] G. F. Utami, D. Suhaedi, and E. Kurniati, "PERBANDINGAN METODE REGRESI LINIER DAN NON-LINIER KUADRATIK PADA PERAMALAN PENJUALAN AIR MINUM," *Matematika : Jurnal Teori dan Terapan Matematika*, vol. 20, no. 2, pp. 33–41, 2021.

[14] K. Ramadani, S. Wahyuningsih, and M. N. Hayati, "FORECASTING STOCK PRICE PT. TELKOM USING HYBRID TIME SERIES REGRESSION LINEAR– AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODEL," *JMSK*, vol. 18, no. 2, pp. 293–307, 2022, doi: https://doi.org/10.20956/j.v18i2.18837.

[15] V. A. Handayani, W. Reza, and S. Mehmood, "FINDING THE BEST MODEL IN NONLINEAR REGRESSION: USING THE COEFFICIENT OF DETERMINATION," *JV*, vol. 7, no. 1, pp. 67–76, 2023, doi: https://doi.org/10.30812/varian.v7i1.2322.

[16] N. R. Draper and H. Smith, *APPLIED REGRESSION ANALYSIS*, 3rd ed. New York: John Wiley & Sons, Inc, 1998.doi: https://doi.org/10.1002/9781118625590

[17] M. F. Qudratullah, *ANALISIS REGRESI TERAPAN*. Yogyakarta: Penerbit Andi, 2024.

[18] D. N. Gujarati, *BASIC ECONOMETRICS*, 4th ed. New York: McGraw-Hill, 2003.

[19] I. Ghozali, *APLIKASI ANALISIS MULTIVARIATE DENGAN PROGRAM IBM SPSS 25*, 9th ed. Semarang: Universitas Diponegoro, 2018.

[20] T. M. Siregar, E. N. Situngkir, J. J. Purba, M. K. Aulia, and R. N. Ramadhani, "MEMPREDIKSI TINGKAT PENGANGGURAN DI KOTA MEDAN DENGAN MODEL REGRESI NON-LINEAR KUADRATIK," *Innovative : Journal Of Social Science Research*, vol. 3, no. 2, pp. 8661–8670, 2023.

[21] R. Mubarak, *PENGANTAR EKONOMETRIKA*, 1st ed. Pamekasan: Duta Media Publishing, 2021.

[22] J. E. Saputra and W. Febrianti, "APPLICATION OF AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) FOR FORECASTING INFLATION RATE IN INDONESIA," *JMSK*, vol. 21, no. 2, pp. 382–396, 2025, doi: 10.20956/j.v21i2.36609.

[23] W. W. S. Wei, *TIME SERIES ANALYSIS : UNIVARIATE AND MULTIVARIATE METHODS*, 2nd ed. Boston: Pearson Education Inc, 2006.

[24] A. Safitri, N. Salam, and A. S. Lestia, "PERAMALAN ANGKA PERCERAIAN DI KABUPATEN JEMBER PADA TAHUN 2022 MENGGUNAKAN METODE ARIMA," *ragam*, vol. 2, no. 2, p. 25, 2024, doi: https://doi.org/10.20527/ragam.v2i2.11333.

[25] R. Ospina, J. A. M. Gondim, V. Leiva, and C. Castro, "An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil," *Mathematics*, vol. 11, no. 14, p. 3069, 2023, doi: 10.3390/math11143069.

[26] W. G. S. Parwita and N. K. N. P. Dewi, "METODE AUTOREGRESSIVE INTEGRATE MOVING AVERAGE DALAM PERAMALAN INDEKS HARGA KONSUMEN KOTA DENPASAR," *SmartEdu Journal*, vol. 1, no. 4, pp. 158–170, 2022.

[27] R. S. Tsay, *ANALYSIS OF FINANCIAL TIME SERIES*. New Jersey: John Wiley & Sons, Inc, 2010.doi: https://doi.org/10.1002/9780470644560

[28] Q. D. H. B. Sitepu, S. Sutarman, and M. A. P. Siregar, "METODE AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) DALAM MEMPREDIKSI JUMLAH PENUMPANG KERETA API KOTA BINJAI," *Arjuna*, vol. 2, no. 2, pp. 69–85, 2024, doi: https://doi.org/10.61132/arjuna.v2i2.621.

[29] R. J. Hyndman and G. Athanasopoulos, *FORECASTING : PRINCIPLES AND PRACTICE*, 2nd ed. Melbourne: OTexts, 2018.doi: https://doi.org/10.32614/CRAN.package.fpp2