Predicting Air Pollution Levels in Jakarta Using Vector Autoregressive Analysis

Khaerun Nisa SH[1], Irfan Irfani[2], Utriweni Mukhaiyar[3]

[1] Master Program in Actuarial, Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology, Jl. Ganesa 10 Bandung 40132, Indonesia.
[2] Master Program in Mathematics, Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology, Jl. Ganesa 10 Bandung 40132, Indonesia.
[3] Statistics Research Division, Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology, Jl. Ganesa 10 Bandung 40132, Indonesia.
utriweni.mukhaiyar@itb.ac.id

Abstract. The air quality in Jakarta is a critical issue affecting public health and the environment. High levels of air pollution can lead to various health problems, including respiratory issues, cardiovascular diseases, and other health disorders. This study aims to predict air pollution levels in Jakarta using Vector Autoregressive (VAR) analysis method on time series data of air pollution levels (AQI) and Particulate Matter (PM2.5) concentrations. VAR is a statistical model used to analyze and forecast time series data consisting of multiple interrelated variables. The research data utilized comprises daily data from IQAir website regarding the air quality index in Jakarta, spanning from August 16 to October 1, 2023. The stages in VAR model analysis consist of: (1) stationary checking and selection model, (2) parameter estimation for selected VAR model, (3) diagnostic checking with normality test, (4) model validation to evaluate the best model, (5) forecasting air pollution levels. The result of research employed VAR(2) model to predict air pollution levels in Jakarta. The VAR(2) model demonstrated stationary data, significant parameter estimates, and relatively small prediction errors, making it the most suitable choice for forecasting air pollution levels. This research will assist the government, environmental organizations, and the public in taking appropriate actions to address high levels of air pollution and protect public health. This study has significant implications for the development of more effective and sustainable air pollution control strategies in Jakarta.

Keywords: Predicting, Air Pollution, Vector Autoregressive.

1 Introduction

The air quality in Jakarta has become a serious issue affecting public health and the urban environment. Previous research has substantiated the adverse cardiovascular and respiratory health impacts resulting from various air pollutants [1]. It is estimated that air pollution has caused 9,600 deaths and incurred approximately $2,500,000,000 USD in losses in Jakarta in 2023 [2]. Two variables are employed for assessing air pollution

N. Djam'an et al. (eds.), *Proceedings of the 5th International Conference on Statistics, Mathematics, Teaching, and Research 2023 (ICSMTR 2023)*, Advances in Computer Science Research 109,
https://doi.org/10.2991/978-94-6463-332-0_3

levels, which are the Air Quality Index (AQI) and PM2.5 air pollution. AQI was developed as a valuable tool for providing information to the community regarding the quality of the air in their vicinity. On the other hand, PM2.5 particles represent a primary component of air pollution that exerts a significant impact on human health. Monitoring the levels of PM2.5 in the air is crucial to protect public health and identify sources of air pollution that need to be addressed.

Forecasting involves making predictions about future occurrences to enable the authorities in the Jakarta region to effectively address this impact. Multivariate time series forecasting involves the utilization of multiple criteria or variables that vary over time [3]. The Vector Autoregressive (VAR) model was chosen because it is one of the multivariate analyses for time series data and allows for examining the interrelationships among variables [4]. It extends the Autoregressive (AR) model by incorporating multiple endogenous variables into the analysis [5]. To mitigate subjective settings in equation model errors, the VAR model considers all variables as endogenous. The VAR model offers several advantages over the traditional single equation model: (1) the VAR model's generality, which allows for the straightforward addition of explanatory variables without strict theoretical constraints; and (2) The VAR model elucidates both short-term and long-term relationships among air quality factors [6].

Additionally, research has investigated the health impacts of air pollution on the population, highlighting the need for effective air quality management [15]. However, there remains a notable research gap when it comes to the utilization of multivariate time series forecasting models, such as the VAR model, in the context of Jakarta air quality. The existing literature often relies on simpler single equation models, which may not fully capture the complex inter-relationships among the various air quality factors in the region. This research aims to offer enhanced predictions of future air pollution levels in Jakarta for the development of more effective and sustainable air pollution control strategies in Jakarta.

2 Vector Autoregressive Model

The VAR model can be estimated without the need to focus on exogenous problems, because all variables are treated as endogenous [7]. In general, the VAR(p) model is defined as follows.

$$Y_t = a + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \quad (1)$$

where Y_{t-p} is vector of size $N \times L$ at times $t-p$, a is vector of size $N \times L$ constants, ϕ_p is coefficient matrix of size $N \times N$ for p, ε_t is residual vector of size $N \times L$, with $(\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{nt})^T$, p is the VAR lag, and t is the observed period.

In the modeling, there is a sequence of steps that must be adhered to when employing vector autoregressive model. This is illustrated as a flowchart in Figure 1.

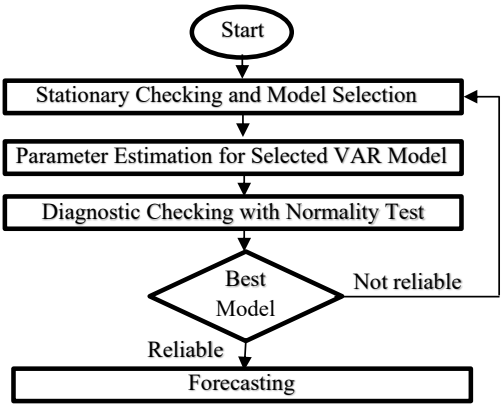

Figure 1. Flowchart of vector autoregressive stages

Based on Flowchart in Figure 1, stages in the VAR model are described as follows.

a. Time series data is considered stationary when both the variance and mean remain constant over time. Stationary with respect to the mean can be formally tested using the Augmented Dickey-Fuller (ADF) test. Additionally, one way to check for stationarity is to examine the eigenvalues of the autocorrelation matrix. If all of the eigenvalues are less than one, then it indicates that the data is likely stationary in mean or constant over time [8].
b. The VAR model selection by systematically testing different orders and selecting the model with the lowest Akaike Information Criterion (AIC), Hannan-Quinn (HQ), Schwarz Criterion (SC), and Final Prediction Error (FPE). These criteria are used to select the appropriate lag order for the VAR model by comparing the goodness-of-fit and model complexity [9].
c. Significance tests for parameters estimation are utilized to identify which parameters significantly influence the model. The t-test can be employed to assess the significance of parameters in the VAR model [10]. This helps researchers identify which parameters are essential for their VAR model and which can be omitted to simplify the model while maintaining its predictive accuracy. Additionally, the normality test of residuals is used to check whether the residuals from the VAR model follow a normal distribution, such as the JB-Test. The JB-Test used to determine whether the null hypothesis can be accepted or reject-ed in a normal distribution [13].
d. Model validation help evaluate the performance of VAR and provide the accuracy of the forecasts, such as RMSE, MAE, and MAPE [14]. Smaller values suggest superior predictive the best model. For a sample of N observations y $(y_i, i = 1,2, \dots, N)$ and N corresponding model prediction $\hat{y}$ with the formula as follows.
   1) Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

   2) Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

   3) Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}}{y_i}\right|$$

3 Result and Discussion

The data employed in this research obtained from the air quality dataset in Jakarta. The dataset spans from August 16 to September 25, 2023, representing t training data, and from September 25 to October 1, 2023, representing the testing data. The results of descriptive analysis of air pollution data are explained as follows.

Table 1. Descriptive Analysis of Pollutant Data in DKI Jakarta Province 2023

Variables	Time	Mean	Std.deviation	Minimum	Maximum

PM2.5	47	55.246	10.098	35	77.7
AQI	47	142.553	15.935	99	162

Multivariate time series analysis is an analytical approach used to model time series data involving multiple interdependent variables. The data utilized in this context consists of pollutant levels, specifically PM2.5 and AQI, as illustrated in Figure 2.

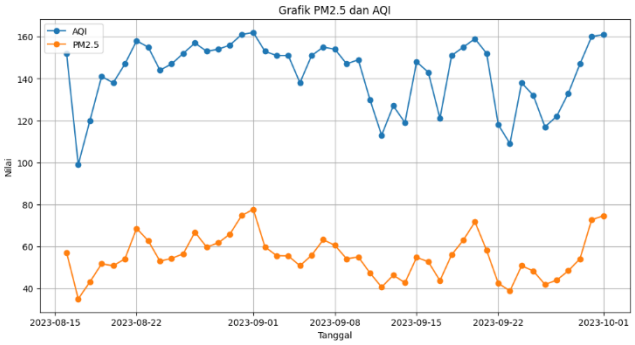

Figure 2. Multivariate Time Series Plot on PM2.5 and AQI Variables

Based on Figure2, the result of PM2.5 and AQI tend to have the same time series pattern with the sample correlation calculation of 0.92. Furthermore, identify the lag model through the Matrix Autocorellation Function (MACF) and Matrix Partial Autocorellation Function (MPACF) plots.

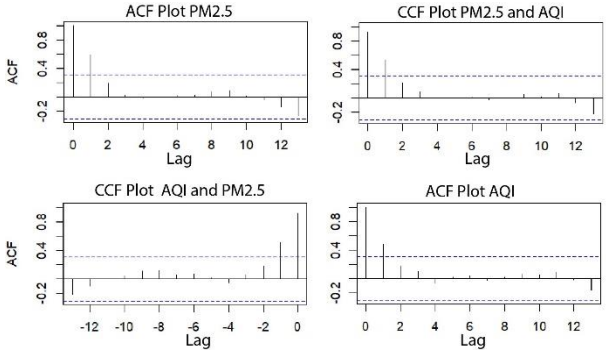

Figure 3. Empiric Plot MACF PM2.5 and AQI

From the MACF sample estimates obtained results as in Figure 3 for the main diagonal reference Autocorrelation Function (ACF) of each variable, whereas for the other components is Cross Corelation Function (CCF). On an empirical MACF model results have an overall result that is tail-off.

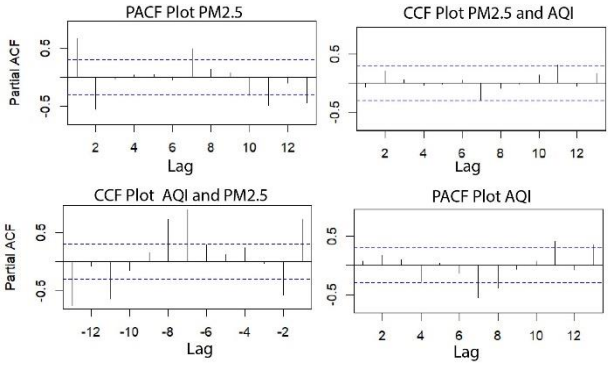

Figure 4. MPACF Empiris PM2.5 dan AQI

In addition to using MACF, the MPACF model is also one of the most commonly used indicators for finding candidates for a multivariate time row model. The results obtained are shown and displayed as in figure 4.3, showing that the MPACF model has a cut-off pattern. Thus, by considering MACF and MPACF, then the candidate that can be chosen for the multivariate time series is the VAR.

Assessment of stationary through mean and variance using ADF test. Based on the results of the ADF test presented in Table 2.

Table 2. The result of ADF test for stationary checking

Variables	P-Value
PM2.5	0.222
AQI	0.078

Based on the ADF test results with a significance level of α is 5%, it can be seen that the PM2.5 and AQI data used are non-stationary data. If the data lacks stationary, the eigenvalues is applied to achieve stationary.

The VAR(p) model fit test was conducted to examine the lag parameters of the VAR(p) model, including AIC(n), HQ(n), SC(n), and FPE(n). The outcomes of the lag parameter assessment for the VAR(p) model are presented in Table 3.

Table 3. Checking lag parameters of the VAR(p) model

Lag	AIC(n)	HQ(n)	SC(n)	FPE(n)
1	8.100	8.184	8.390	3303.111
2	8.222	8.362	8.706	3761.557
3	8.511	8.706	9.189	5110.612

Considering the computed results of the four selection criteria, despite VAR(1) having the lowest values for AIC(n), HQ(n), SC(n), and FPE(n), but based on the p-value for the estimated parameter VAR(1) with the p-value $\phi^{(1)}_{PM2.5:PM2.5}$ is 0.055 and VAR(2) with the p-value for $\phi^{(1)}_{PM2.5:PM2.5}$ is 0.023. This indicates that the VAR(1) model lacks significant parameter estimates. Therefore, VAR(2) was chosen because it has the next lowest values for AIC(n), HQ(n), SC(n), and FPE(n) after VAR(1), and it also satisfies the significance criteria for parameter estimation values.

After identifying the suitable VAR model, the subsequent stage involves estimating the model parameter using techniques, such as the t-test. VAR with lag-parameters p and m time series variables is defined as:

$$\begin{pmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{mt} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} + \begin{pmatrix} \phi_{11}^{(1)} & \cdots & \phi_{1m}^{(1)} \\ \phi_{21}^{(1)} & \cdots & \phi_{2m}^{(1)} \\ \vdots & \ddots & \vdots \\ \phi_{m1}^{(1)} & \cdots & \phi_{mm}^{(1)} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{2,t-1} \end{pmatrix} + \cdots + \begin{pmatrix} \phi_{11}^{(p)} & \cdots & \phi_{1m}^{(p)} \\ \phi_{21}^{(p)} & \cdots & \phi_{2m}^{(p)} \\ \vdots & \ddots & \vdots \\ \phi_{m1}^{(p)} & \cdots & \phi_{mm}^{(p)} \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{2,t-p} \end{pmatrix}$$

Estimation of the VAR(2) model parameters using the smallest square method obtained the following result:

$$\begin{pmatrix} Y_{1t} \\ Y_{2t} \end{pmatrix} = \begin{pmatrix} 17.815 \\ 81.671 \end{pmatrix} + \begin{pmatrix} 0.785 & -0.022 \\ 0.743 & 0.216 \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \begin{pmatrix} -0.536 & 0.191 \\ -0.543 & 0.141 \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix}$$

with the eigenvalues is $\max_{1 \le i \le 4} \{|\lambda_i|\} = 0.669 < 1$. Based on the eigenvalue principle in the VAR model, the data used is stationary [8].

Furthermore, to check whether the residual of the VAR model follow a normal distribution, the JB-Test is employed on multivariate time series data. Based on the results of the normality test presented in Table 4.

Table 4. Residual Normality Test Results

Uji	$\chi^2 - value$	$p - value$
JB-Test	4.748	0.314

Obtained residual vector results is multivariate normal distribution, then using the VAR(2) model will be made predictions along out-sample data, with a lot of test data is 6 days. With error results prediction results and out-sample data measured results obtained in Table 5.

Table 5. VAR(p) Model Error Calculation Result and Out-Sample Data

Model	Galat	PM2.5	AQI	Model	Galat	PM2.5	AQI
VAR (1)	RMSE	25.986	60.880	VAR (6)	RMSE	25.630	60.748
	MAPE	0.380	0.331		MAPE	0.392	0.338
	MAE	21.231	46.330		MAE	21.883	47.142
VAR (2)	RMSE	25.925	61.660	VAR (7)	RMSE	25.082	60.657
	MAPE	**0.379**	**0.325**		MAPE	0.383	0.333
	MAE	**21.078**	**45.401**		MAE	21.928	46.502
VAR (3)	RMSE	25.720	61.048	VAR (8)	RMSE	25.018	59.962
	MAPE	0.386	0.325		MAPE	0.380	0.333
	MAE	21.089	45.469		MAE	21.285	46.485
VAR (4)	RMSE	25.596	61.127	VAR (9)	RMSE	**24.607**	59.202
	MAPE	0.384	0.329		MAPE	0.381	0.333
	MAE	21.368	45.960		MAE	21.168	46.464
VAR (5)	RMSE	25.877	60.928	VAR (10)	RMSE	25.746	**58.709**
	MAPE	0.391	0.333		MAPE	0.404	0.343
	MAE	21.765	46.545		MAE	22.464	47.856

In accordance with the information provided in the table, it is apparent that the VAR(9) and VAR(10) models have the smallest errors according to the RMSE calculations, whereas VAR(2) demonstrates the four smallest errors when considering MAPE and MAE calculations. Thus, the selected model for validation is VAR(2). The following are the forecasting results of the VAR(2) model from the entire data in Table 6.

Table 6. Forecasting the air pollution levels with VAR(2) model

No	Prediction PM2.5	Real Data PM2.5	Prediction AQI	Real Data AQI
1	64.335	55.6	154.827	151
2	55.225	51.8	145.167	141
3	52.742	55.8	140.76	151
4	54.22	53.2	141.498	144

Predictions based on the VAR(2) model tend to decrease for PM2.5 and AQI values. When compared with the original data, the data shows a similar trend, indicating that the selected model provides accurate predictions, spanning from October 2 to October 5, 2023. The following is a forecast graph from the VAR(2) model for the next four days.

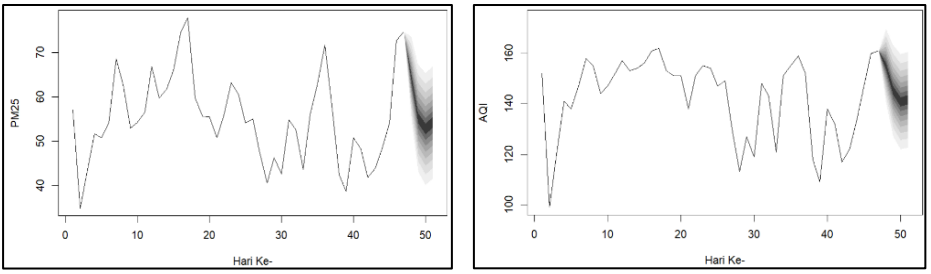

Figure 5. VAR(2) model prediction results from all data

4 Conclusion

VAR models are well-suited for analyzing and forecasting time series data with multiple interconnected variables, making them particularly relevant for air pollution prediction. Air pollution levels can be influenced by a wide range of external factors, including meteorological conditions, industrial activities, and government policies. These external factors may not have been fully accounted for in the research. Additionally, future research can build upon this work by exploring more considering external factors, and broadening the scope of Spatial and Temporal analysis to improve air quality forecasting. This research aimed to predict air pollution levels in Jakarta using VAR modeling, which allows the analysis and forecasting of time series data with multiple interconnected variables. From the data analysis and exploration, it can be concluded that the stationarity requirement is met with a maximum eigenvalue of 0.669. Based on the parameter estimation, VAR(2) has significant parameter estimates. Furthermore, VAR(2) exhibits relatively smaller MAPE and MAE values compared to the other models, even though it's not the smallest in terms of RMSE. However, following the principle of model parsimony, VAR(2) is considered the best model. As for the predictions

of the VAR(2) model, it tends to decrease and shows a similar trend when compared with the original data. In summary, this study successfully employed VAR(2) modeling to predict air pollution levels in Jakarta. The selected model demonstrated stationarity, significant parameter estimates, and relatively small prediction errors, making it the most suitable choice for forecasting air pollution levels.

References

1. Afgun, U.R.S., et al.: A spatial feature engineering algorithm for creating air pollution health datasets. International journal of cognitive computing in engineering, vol. 1, 98-107 (2020).
2. IQAir Hompage, https://www.iqair.com/id/indonesia/jakarta last accessed 2023/9/17.
3. Zahara, S., and Sugianto.: Forecasting Consumer Index Data Based on Multivariate Time Series Using Deep Learning. Journal RESTI 5(1), 24-30 (2021).
4. Desviana, A.P., and Maryam, J.: Air Pollution Modeling Using the VAR Method in Riau Province. Journal of Science, Technology and Industry 13(2), 160-167 (2016).
5. Arianto, A.T., Parmikanti, K., Suhandi, B., and Ruchjana, B.N.: Forecasting Particulate Matter 2.5 (PM2.5) Concentrations using the Vector Autoregressive Model with the Maximum Likelihood Estimation Method. Journal KUBIK 6(1), 1-12 (2021).
6. Aasim, Singh, S.N., and Mohapatra, A.: Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting. Renewable energy, vol. 136, 758-768 (2019).
7. Lütkepohl, H.: Econometric Analysis with Vector Autoregressive Models. Departement of Economics European University Institute, Italy (2007).
8. Wei, W.W.S.: Time Series Analysis Univariate and Multivariate Methods. 2nd edition. Pearson Addison-Wesley, New York (2006).
9. Kirchgassner, G., and Wolters, J.: Introduction to Modern Time Series Analysis. Spinger, Berlin (2007).
10. Montgomery, D.C., Jennings, C.L., and Kulachi, M.: Introduction to Time Series Analysis and Forecasting. John Wiley & Sons, New Jersey (2015).
11. Mukhaiyar, U., Widyanti, D., and Vantika, S. The time series regression analysis in evaluating the economic impact of COVID-19 cases in Indonesia. Journal Model Assisted Statistics and Applications 16(3), 197-210 (2021).
12. Yuliati, I.F., Istinah, A.N., and Sihombing, P.R.: Application of Vector Autoregressive Integrated (VARI) on Data on the Number of Active Family Planning Participants. Journal Mathematics and Applied Scientific 17(2), 258-272 (2020).
13. Jarque, C.M., and Bera, A.K.: Efficient Tests for Normality, Homoskedasticity, and Serial Independence of Regression Residual. Economics Letters, vol. 6, 255-259 (1980).
14. Dai, H., Huang, G., Wang, J., and Zeng, H.: VAR-tree model based spatio-temporal characterization and prediction of O_3 concentration in China. Ecotoxicology and Environmental Safety, 257, 1-15 (2023).
15. Dominski, F.H., Branco, J.H, Buonanno, G., and Stabile, H.: Effects of air pollution on health: A mapping review of systematic reviews and meta-analyses. Environmental Research, 201, 111487 (2021)