Air Quality Forecasting in Surabaya using VAR and VARX

Nadya Yuniar Desi Prameswari[1], Adatul Mukarromah[2]
Department of Statistics, Faculty of Science and Data Analytics,
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: yuniarnadya4@gmail.com(1), adatul@statistika.its.ac.id(2)

***Abstract*—Surabaya is the center of industry and trade in East Java. As the capital of East Java Province and metropolitan city, Surabaya has experienced increasing population and vehicles volume from year to year. The increasing crowd of Surabaya results in the emergence of various problems, one of which is regarding air pollution. Air pollution also can lead to the issue regarding health. Based on the problems caused by air pollution, the quality of the air needs to be considered. The air quality can be seen through the Air Pollutant Standard Index (ISPU) obtained through AQMS at the SUF station, with several indicators used to determine air quality, namely CO, PM_{10}, and NO_2. One of the preventive actions that can monitor air quality is forecasting the levels of indicators that determine air quality. So this study aims to predict CO, PM_{10}, and NO_2 at SUF 7 station Surabaya using VAR and VARX. The analysis results show that the three indicators have a daily seasonal pattern per half hour. The total levels of the three indicators tend to be low in the afternoon to evening and high in the morning to noon. In the VAR model formed, diagnosing checking for the assumption of residual white noise and multivariate normal distribution is not met. These assumptions have not been met due to the outliers in the data used. In addition, the results show that the VARX method in train data produced the best model for CO, PM_{10}, and NO_2 indicators.**

***Keywords*— Air Quality, Forecasting, Time Series, VAR, VARX.**

I. INTRODUCTION

Surabaya is the capital of East Java Province. Surabaya is also the center of industry and trade in East Java. As a metropolitan city, the traffic volume in Surabaya is high. The number of vehicles in Surabaya has continuously increased quite sharply [1]. The increasing number of vehicles in Surabaya is caused by the city's growth and its population. The increasingly crowded in Surabaya results in various problems, one of which is regarding air pollution.

One of the leading causes of air pollution in Surabaya is the transportation sector, where the transportation sector produces 60 percent of carbon monoxide (CO) [2], increasing the amount of particulate matter (PM_{10}) [3] and NO_2 in the air, which can give negative effects on the environment and human health [4]. Pollutants that cause air pollution are estimated to have caused 350 deaths in Surabaya by 2021 [5]. Based on this, the level of air pollution in Surabaya has reached an alarming point and has become a problem that must be resolved immediately.

The air quality also directly or indirectly affects the lives of living things. Hence, the air quality needs to be controlled. One of the preventive actions that can monitor air quality is by forecasting pollutants that determine air quality. Forecasting of air quality has been widely carried out, some of which are using ARIMA [6] [7], DSARIMA [8] [9], FFNN [7], MGSTAR [6] [9], and LSTM [7]. Some of these studies have predicted air pollutants univariately. Whereas existing air pollutants can form a mixture that allows linkages between air pollutants, a modeling method is needed to capture differences in air pollutant parameters. In this study, it is suspected that there is a correlation between air pollutants that determine air quality. Based on this, this study will forecast air pollutants CO, PM_{10}, and NO_2 at the SUF 7 station using Vector Autoregressive Moving Average (VAR) and Vector Autoregressive with Exogenous Variables (VARX). This study chose SUF 7 station because the permanent monitoring station (SUF) 7 is located in the Kebonsari area, which is an area with several schools, residential areas, and crowded packed offices/companies. In addition, there are also several main highways, and traserved by the Surabaya-Porong toll road so that this area is almost always crowded with traffic activities. The VAR method is used because this model can explain the linear relationship between observations with different variables, which can be seen through the Cross-Correlation Function or CCF [10]. In addition, time-series data is often influenced by an event that causes the residuals assumption multivariate normal distribution to be not met. Hence, this study will use VARX as another method to forecast air pollutants. Furthermore, this study will select the best model through the smallest Root Mean Square Error (RMSE). The best model is then used to forecast the air pollutant determining the ISPU used at the SUF 7 station. Hence it can provide information about the value of the air pollutant determining the ISPU and give consideration in tackling air pollution to the Surabaya City Environmental Department or the Surabaya Government.

II. LITERATURE REVIEW

A. Expectation-Maximization (EM)

Expectation-Maximization is an imputation method based on least squares and maximum likelihood estimation [11]. The maximum likelihood uses all available data, both complete and incomplete, to obtain the estimated parameter values with the highest chance of producing sample data. EM consists of iterative calculations, namely prediction and estimation.

B. Correlation Analysis

Correlation analysis is a method used to determine whether or not there is a relationship between one variable and another. This method has a value that varies from -1 to 1, where -1 means a minus total correlation, 0 indicates no correlation, and 1 shows a perfect positive correlation. [12].

$$r_{Z_1 Z_2} = \frac{\sum_{i=1}^{n}(Z_1 - \bar{Z}_1)\,(Z_2 - \bar{Z}_2)}{\sqrt{\sum_{t=1}^{n}(Z_1 - \bar{Z}_1)^2}\sqrt{\sum_{i=1}^{n}(Z_2 - \bar{Z}_2)^2}} \quad (1)$$

where

Z_1 : 1st variable
Z_2 : 2nd variable
n : the amount of the data
r : correlation coefficient with terms $-1 < r < 1$

The hypothesis in this test is as follows.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

The test statistics used are as follows.

$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad (2)$$

H_0 will be rejected if $t_{calculated} > t_{\frac{\alpha}{2},n-2}$ or if $t_{calculated} < -t_{\frac{\alpha}{2},n-2}$.

C. Vector Autoregressive (VAR)

Several AR models form the *Vector Autoregressive* (VAR) model, where AR models create a vector whose variables influence each other. Equation 3 contains the general form of the AR model (*p*) [13].

$$Z_t = \phi_1 Z_{t-1} + \cdots + \phi_p Z_{t-p} + a_t \quad (3)$$

where

Z_t $= Z_t - \mu$,
a_t : residual at time-t^{th},
p : order AR,
$\phi_p(B)$ $= 1 - \phi_1 B - \phi_2 B^2 - \cdots \phi_p B^p$

$\phi_1, \phi_2, \cdots \phi_p$: AR coefficient of order *p*.

A multivariate time series can be a VAR model equation with the order *p* or VAR(*p*) if it follows the following equation [14].

$$\dot{\boldsymbol{Z}}_t = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_1 \dot{\boldsymbol{Z}}_{t-1} + \boldsymbol{\Phi}_p \dot{\boldsymbol{Z}}_{t-p}\mathbf{a}_t \qquad p > 0, \quad (4)$$

where

$\dot{\boldsymbol{Z}}_{t-p}$: vector $m \times 1$ of the variable at time-*(t-p)*th
$\boldsymbol{\Phi}_p$: matrix $m \times m$ of parameter p^{th}
$\mathbf{a}_t$: vector of residuals at time-t^{th}

1. Stasionery of Data

Non-stationary data are often encountered in time series analysis. Overcoming non-stationarity in the variance can use the Box-Cox transformation. Equation 5 contains The Box-Cox transformation formula.

$$T(Z_t) = \begin{cases} \dfrac{Z_t^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \lim\limits_{\lambda \to 0} \dfrac{Z_t^{\lambda} - 1}{\lambda} = \ln Z_t, & \lambda = 0 \end{cases} \quad (5)$$

where λ is the parameter of the Box-Cox transformation. Augmented Dicky Fuller is one of the methods of testing the stationarity of the data in the mean. This test is based on the following model equation.

$$Z_t = \phi Z_{t-1} + a_t \quad (6)$$

where a_t is a residual *white noise*. The tests carried out are as follows.

$H_0 : \phi = 1$
$H_1 : \phi < 1$

The test statistics used are as follows.

$$DF = t\ ratio = \frac{\hat{\phi} - 1}{std(\hat{\phi})} = \frac{\sum_{t=1}^{n} Z_{t-1} + a_t}{\hat{\sigma}\sqrt{\sum_{t=1}^{n} Z_{t-1}^2}} \quad (7)$$

at a significance level 5%, $\hat{\phi}$ and $\hat{\sigma}$ obtained from the estimation results of the method least square.

2. Matrix Cross Correlation Function (MCCF)

For example, if there is a vector *time series* with observations *n*, namely $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$, the sample correlation matrix equation is as follows.

$$\hat{\rho}(k) = \left[\hat{\rho}_{ij}(k)\right] \quad (8)$$

with $\hat{\rho}_{ij}(k)$ is a cross-correlation of samples for component series to-*i* and to-*j*.

3. Matrix Partial Cross Correlation Function (MPCCF)

Heyse and Wei obtained the equation for the autocorrelation matrix lag partial on lag s, the equation is as follows [13].

$$\boldsymbol{P}(s) = [\boldsymbol{D_v}(s)]^{-1}\boldsymbol{V_{vu}}(s)[\boldsymbol{D_u}(s)]^{-1} \quad (9)$$

where

$\boldsymbol{D_v}(s)$: the diagonal matrix where the element-i^{th} is the root of the diagonal element-i^{th} of $\boldsymbol{V_v}(s)$
$\boldsymbol{D_u}(s)$: the diagonal matrix where the element-i^{th} is the root of the diagonal element-i^{th} of $\boldsymbol{V_u}(s)$.

4. Identification of VAR Model

The VAR model can be identified by looking at *time series* plots, MCCF plots, and MPCCF plots to determine the model's order of *p* and *q*. In addition, determining the order of VAR can also be done through *Akaike's Information Criterion* (AIC). The AIC formula is as follows.

$$AIC_{(p+q)} = \ln\left|\hat{\boldsymbol{\Sigma}}_{(p+q)}\right| + \frac{2m^2(p+q)}{n} \quad (10)$$

where

$\hat{\boldsymbol{\Sigma}}_{(p+q)}$: estimation of the matrix variance-covariance
p : order AR

q : order MA
$2m^2(p+q)$: number of parameters of AR and MA
n : the amount of data

5. *Parameter Estimation*

After the temporary model has been formed, then the parameter estimation of the VAR model is carried out using LS (*Least Square).* For example, there is a VAR model as follows.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U} \tag{11}$$

where

$\boldsymbol{Y}$ = $(\boldsymbol{Z_1}, \boldsymbol{Z_2}, \dots, \boldsymbol{Z_n})$
$\boldsymbol{B}$ = $[\boldsymbol{\Phi} = \boldsymbol{\Theta}]$
$\boldsymbol{U}$ = $(\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_n})$
$\boldsymbol{y}$ = **Vec(Y)**
$\boldsymbol{\beta}$ = **Vec(B)**
$\boldsymbol{u}$ = **Vec(U)**

6. *Parameter Significance Test*

A parameter significance test was conducted to determine the significant parameters of the model using the t-test. The test hypothesis used is as follows.

$H_0 : \emptyset_{ijk} = 0$
$H_1 : \emptyset_{ijk} \neq 0$

The test statistic used is written in Equation 12.

$$t = \frac{\widehat{\emptyset}_{ijk}}{SE\left(\widehat{\emptyset}_{ijk}\right)} \tag{12}$$

H_0 will be rejected if $|t_{calculated}| > t_{\frac{\alpha}{2},n-p-1}$ or if the value of the *p-value* is less than α, which p shows the number of parameters means that the parameter has been significant.

7. *Diagnostic Checks*

VAR models with significant parameters must meet the assumptions of white noise and distribution multivariate normal. The hypothesis to test the assumption *white noise* using the Portmanteau test is as follows.

H_0 : model residual vector meets assumption white noise
H_1 : model residual vector does not meet assumption white noise

The test statistic used is written in Equation 13.

$$Q_h = n\sum_{i=1}^{h} tr\left(\hat{C}_i'\hat{C}_0^{-1}\hat{C}_i\hat{C}_0^{-1}\right) \tag{13}$$

C_i obtained from $\hat{C}_i = n^{-1}\sum_{t=i+1}^{n} \hat{a}_t\hat{a}'_{t-i}$.

H_0 will be rejected if $Q_h > X^2$ or if the *p-value* is less than α [15]. Furthermore, check whether the residuals have met the assumption of a multivariate distribution normal carried out by performing the Shapiro-Wilk test. The hypothesis to test is as follows.

H_0 : the residual vector of the model meets the assumption of a distribution multivariate normal
H_1 : the residual vector distribution model does not meet the assumption of multivariate normal.

The test statistic used is written in Equation 14.

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})} \tag{14}$$

where

$\boldsymbol{Y}$ = $(\boldsymbol{Z_1}, \boldsymbol{Z_2}, \dots, \boldsymbol{Z_n})$
$\boldsymbol{B}$ = $[\boldsymbol{\Phi}: \boldsymbol{\Theta}]$
$\boldsymbol{U}$ = $(\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_n})$
$\boldsymbol{y}$ = **Vec(Y)**

Failed to reject H_0 if the p-value is greater than α [16].

8. *Model Goodness Criteria*

One of the criteria for the value *error* smallest in the data *train* and data *test* that can be used to select the best model is Root Mean Square Error (RMSE). The formula for calculating RMSE is as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Z_{i,t} - \hat{Z}_{i,t}\right)^2} \tag{15}$$

where $Z_{i,t}$ represents real value and $\hat{Z}_{i,t}$ represents forecast.

D. Residual Control Diagram

External events often affect data time series. If the cause of this is not known, it is called an outlier. However, if the cause of the incident is known, it is called an intervention. The residual control chart is one way that can be used to detect the presence of outliers is performed by the process control targets and classification of individual observations using T^2 Hotelling [17].

$$\boldsymbol{J_{i,h}} = \boldsymbol{\omega'_{i,h}\Sigma^{-1}_{i,h}\omega_{i,h}} \tag{16}$$

where $\boldsymbol{\omega_{i,h}}$ is the residual of each observation in the column vector.

E. Vector Autoregressive with Exogenous Variable (VARX)

The VARX method is a development of the VAR model by adding an exogenous variable to the right of the equation. Variable X as an exogenous variable is included in the VAR model so that the model of VARX (p,s) is as follows [18].

$$\boldsymbol{\Phi_p}(B)\boldsymbol{Z_t} = \gamma_r(B)\boldsymbol{X_t} \tag{17}$$

where

$\boldsymbol{\Phi_p}(B)$ = $\mathbf{I_k} - \Phi_1 B - \cdots - \Phi_R B^p$
$\gamma_r(B)$ = $\gamma_0 + \gamma_1 B - \cdots - \gamma_r B^r$

$\boldsymbol{\Phi_p}$ is a matrix of size $(m \times m)$, and γ_r is a matrix of size $(m \times r)$.

By modeling VARX $\boldsymbol{Z_i} = X\beta + D\delta + \varepsilon$, and parameter estimates obtained for δ and β are as follows [19].

$\hat{\delta}$ = $D^{-1}(1-M)(Z - X[X^T\Sigma^{-1}MX]^{-1}[X^T\Sigma^{-1}MZ])$
$\hat{\beta}$ = $[X^T\Sigma^{-1}MX]^{-1}[X^T\Sigma^{-1}MZ]$

which $\hat{\beta}$ is an estimator for the VARX model, $\hat{\delta}$ is an estimator for *exogenous variables*, and $M = [I - D(D^T\Sigma^{-1}D)^{-1}D^T\Sigma^{-1}]$, with the minimum variance requirements, meet normal distribution and unbiased.

F. Air Pollution

Air pollution is pollution caused by human activities such as factories, motor vehicles, burning garbage, agricultural residues, and natural events such as forest fires and volcanic eruptions that emit dust, gas, and hot clouds [20]. The air pollutants used to calculate the Air Pollutant Standard Index (ISPU) are PM_{10}, CO, SO_2, NO_2, and O_3.

III. RESEARCH METHOD

A. Data Source

The data used in this study is secondary data resulting from air quality monitoring at SUF 7 station in Surabaya per half-hour conducted by the Surabaya City Environment Service. September 2020 data is used as *train* data, and data for the first week of October 2020 is used as a data *test*.

B. Research Variables

Research variables used in this study are as follows.

Table 1. Research Variables

Symbol	Variable	Description
Z_1	CO	Carbon Monoxide
Z_2	PM_{10}	Particulate Matter
Z_3	NO_2	Nitrogen Dioxide

C. Data Structure

The data structure of the variables used is as follows.

Table 2. Data Structure

t	Hour	Month	$Z_{1,t}$	$Z_{2,t}$	$Z_{3,t}$
1	0:30	September	$Z_{1,1}$	$Z_{2,1}$	$Z_{3,1}$
2	1:00	September	$Z_{1,2}$	$Z_{2,2}$	$Z_{3,2}$
⋮	⋮	⋮	⋮	⋮	⋮
1776	00:00	October	$Z_{1,1776}$	$Z_{2,1776}$	$Z_{3,1776}$

the index used in this study is $Z_{i,t}$ with:

i : index of the type of ISPU air quality determinant with $i = 1, 2, 3$. 1 is CO content, 2 is PM_{10}, and 3 is NO_2,

t : index of the time sequence of events with $t = 1, 2, \ldots, 1776$.

D. Analysis Steps

The analysis steps are to obtain forecasting results indicators that determine air quality are as follows.

1. Perform preprocessing of data. If there is a missing value shall be made on imputation missing value using Expectation Maximization.
2. Describe the characteristics of data from air quality on indicators of CO, PM_{10}, and NO_2 at SUF 7 station in Surabaya.
3. Perform correlation test.
4. Split data into train data and data test.
5. Perform modeling and forecasting on indicators of CO, PM_{10}, and NO_2 at SUF 7 station in Surabaya using the method *Vector Autoregressive* (VAR) with the following steps.
   (i) Identifying stationarity in the *mean* and *variance of the* three air quality indicators used.
   (ii) Identify orders based on MCCF and MPCCF plots.
   (iii) Estimating model parameters and testing the significance of model parameters.
   (iv) Checking the diagnosis of assumptions *white noise* and having a normal distribution *multivariate* on the residual data used.
   (v) Forecasting air quality parameters CO, PM_{10}, and NO_2 at SUF 7 station in Surabaya using the VAR method.
   (vi) Selection of the best VAR model.
6. Modeling and forecasting the CO, PM_{10}, and NO_2 indicators at SUF 7 station in Surabaya using the *Vector Autoregressive with Exogenous Variables* (VARX) method with the following steps.
   (i) Detect *outliers* using residual control diagrams.
   (ii) Creates variables *dummy* as many as *outliers* found in detection *outlier*
   (iii) Identify orders based on MCCF and MPCCF plots.
   (iv) Estimating model parameters and testing the significance of model parameters.
   (v) Checking the diagnosis of assumptions white noise and having a normal distribution multivariate on the residual data used.
   (vi) Forecasting air quality parameters of CO, PM_{10}, and NO_2 at SUF 7 station in Surabaya using the VARX.
   (vii) Selection of the best VARX models.
7. Forecasting the air quality parameters of CO, PM_{10}, and NO_2 at SUF 7 station in Surabaya based on the best model.
8. Draw conclusions and suggestions.

IV. ANALYSIS AND DISCUSSION

A. Characteristics of Air Quality Determinant Indicators at SUF 7 Station Surabaya

The characteristics of CO, PM_{10}, and NO_2 indicators measured using AQMS at SUF 7 Station located in Kebonsari can be seen in Figure 1.

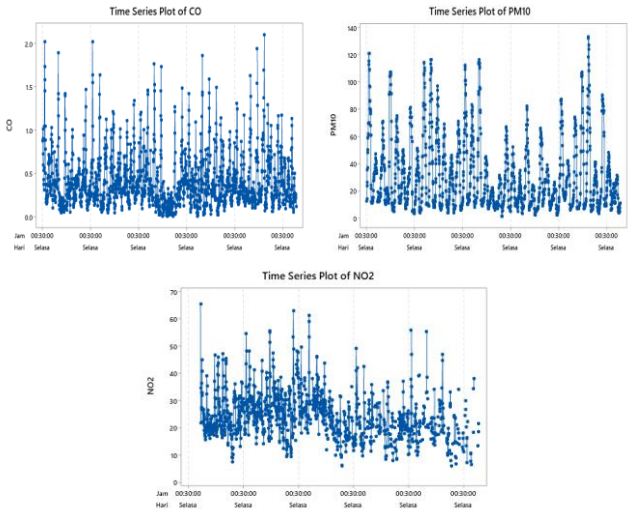

Figure 1. Time Series Plot of CO, PM_{10}, and NO_2 at SUF 7 Station

Figure 1 shows CO, PM_{10}, and NO_2 levels at SUF 7 Station in Surabaya, which is recorded every half hour from September 1st to October 7th 2020. Figure 1 also shows missing data in the data used, so it is necessary to impute the *missing value* before further analysis is carried out.

In this study, the imputation method *missing value* used is *Expectation Maximization*. The characteristics of the data used after the imputation process *missing value* can be seen in Table 3.

Table 3. Characteristics of Determinants of Air Quality

Variable	*Mean*	Minimum	Maximum
CO	0,413343	0,00135	2,1071
PM_{10}	30,0911	0,06666	133
NO_2	24,0575	5,8684	65,551

Table 3 shows that the level of PM_{10} indicator has the highest average level of 30,0911. The level of CO indicator has the smallest minimum value compared to other indicators, namely 0,00135, while the maximum value of the indicator levels of NO_2 reaches 65,551. Then for more details in viewing the pattern of data on air quality determinants, it can be seen through Figure 2.

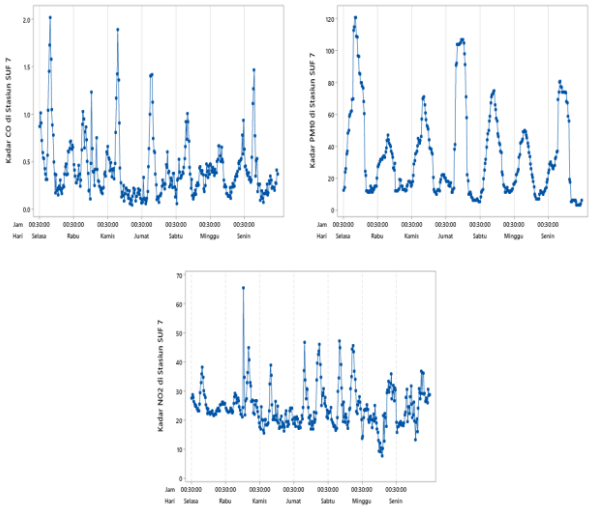

Figure 2. Time Series Plot of CO, PM_{10}, and NO_2 in 7 Days

Based on Figure 2, the three air quality indicators on Sunday tend to have a lower number of levels than the number of levels on other days. In addition, the levels of the three air quality indicators tend to be high at 08.00. A large number of levels at this hour is in line with the start of the activities of the citizens of Surabaya. At 16.00, the levels of the three air quality indicators have low fluctuations that are directly proportional to the end of the citizens of Surabaya. Based on this, the levels of indicators that determine air quality CO, PM_{10}, and NO_2 at SUF 7 Station in Surabaya have a seasonal pattern, namely daily seasonality per half hour.

B. Correlation Analysis

In this study, correlation tests were conducted on the three variables that determine air quality at SUF 7 Station in Surabaya to determine whether there is a relationship between the variables used. The results of correlation testing in this study can be seen in Table 4.

Table 4. Results of Correlation Analysis

Variable	*P-value*	Decision	Correlation Level
CO - PM_{10}	$2{,}46 \times 10^{-30}$	Tolak H_0	0,27
CO - NO_2	$7{,}62 \times 10^{-94}$	Tolak H_0	0,46
PM_{10} - NO_2	$2{,}31 \times 10^{-12}$	Tolak H_0	0,17

Based on Table 4, there is a correlation between the variables used. The level of closeness between the CO indicator-the PM_{10} indicator and the PM_{10}-NO_2 indicators has a weak level of closeness, this is evidenced by the correlation values of 0,27 and 0,17. While the indicators of CO and NO_2 found a correlation value of 0,46 or has a strong relationship. Based on the relationship between these indicators, the analytical methods that will be used for modeling are VAR and VARX.

C. Modeling of Air Quality Determinant Indicators at SUF 7 Station in Surabaya with VAR

The stages carried out in modeling using VAR can be seen in the following discussion sub-sections.

1. *Stasionerity of the Data*

In identifying the stationarity of the data, the data used must be stationary in the *variance* and stationary in the *mean*. The results of the identification of data stationarity can be seen as follows.

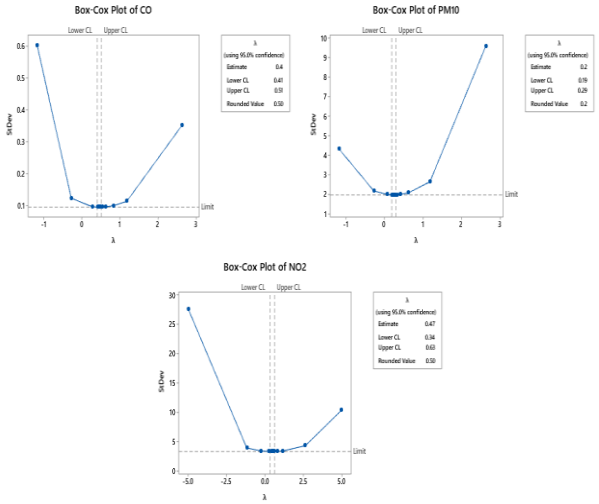

Figure 3. Box-Cox Plot Before Transformation

Based on Figure 3, none of the three indicators that determine air quality is stationary in variance. This can be seen through the *rounded value* of each variable with a value of less than 1,00 so that the next transformation is carried out.

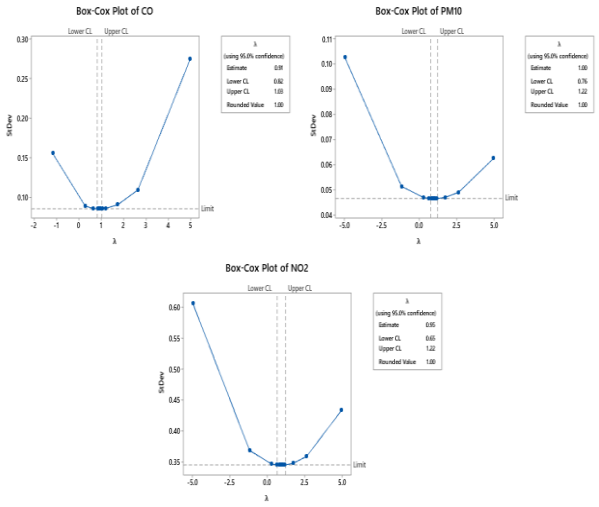

Figure 4. Box-Cox Plot After Transformation

After the transformation process was carried out, the three air quality determinants were stationary in *variance*. So the identification can be carried out to the next stage, namely the identification of the stationarity of the data to the *mean*, which can be done by using the Augmented Dickey-Fuller test.

Table 5. Test Results Augmented Dickey-Fuller Non-Seasonal

Variable	P-value	Conclusion
CO	0,01	Stationary
PM_{10}	0,01	Stationary
NO_2	0,01	Stationary

Based on Table 5, the data used in this study was stationary in the mean. This can be seen through the p-value of each indicator used having a value of less than alpha (0,05). In Figure 2, the data used indicate an alleged seasonal pattern. Seasonal patterns will be identified univariately using the ACF plot.

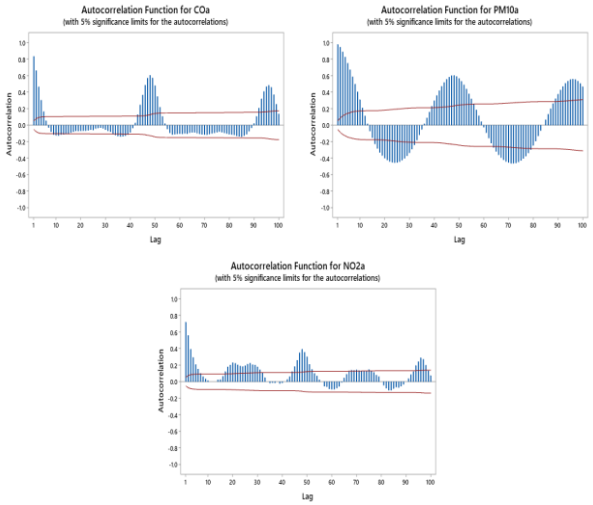

Figure 5. ACF Plot of Air Quality Determinant Indicators at SUF 7 Station Surabaya

Before the *differencing, it* can be seen that after lag 1, the next lag on the three indicators tends to step is to *dies down* and has a seasonal pattern of 48. Since the data has a seasonal pattern of 48, the next carries out *differencing* 1 and *differencing* 48. The seasonal Augmented Dickey-Fuller test results that have been *differencing* can be seen in Table 6.

Table 6. Augmented Dickey-Fuller Test Results Seasonal

Variable	P-value	Conclusion
CO	0,01	Stationary
PM_{10}	0,01	Stationary
NO_2	0,01	Stationary

After *differencing*, the p-value of the seasonal ADF results is less than alpha. This indicates that the data air quality determinant indicators at SUF 7 station in Surabaya are stationary in the *mean* and can be used to analyze the next stage.

2. *Identification of VAR Model*

Identification of the VAR model can be seen through Matrix Partial Cross-Correlation Function (MPCCF) plots. The MPCCF plots in this study are as follows.

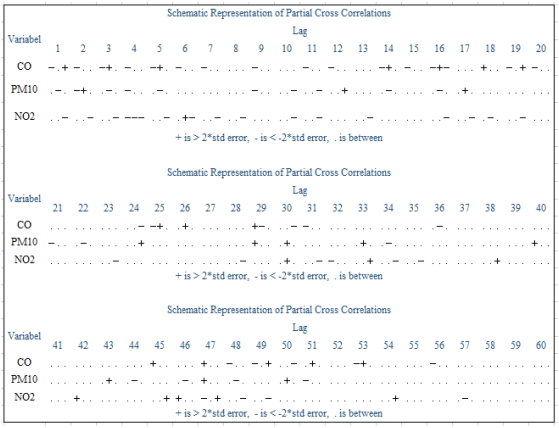

Figure 6. MPCCF Plot of Air Quality Determinant Indicators

The MPCCF plot of air quality determinants at SUF 7 Station in Surabaya is significant in the 1st lag to the 5th lag. While the lag that shows a seasonal pattern is in the 48th lag. So the VAR model formed is $(5,1,0)(1,1,0)^{48}$.

3. *Parameter Estimation and Parameter Significance Test*

Parameter estimation of the VAR model was carried out on the VAR $(5,1,0)(1,1,0)^{48}$ model and found as many as 171 parameters. Furthermore, a significance test was conducted using *restrict* to find out the significant parameters to the model.

Table 7. Results of VAR Model Parameter Estimation $(5,1,0)(1,1,0)^{48}$

Equation	Parameter	Estimasi	Std Error	t-Value	P-Value	Variabel
CO	$\emptyset_{(1,1,1)}$	-0,351	0,025	-13,7	0,0001	$CO_{(t-1)}$
	$\emptyset_{(2,1,1)}$	-0,146	0,023	-6,2	0,0001	$CO_{(t-2)}$
	$\emptyset_{(3,1,1)}$	-0,159	0,023	-6,81	0,0001	$CO_{(t-3)}$
	$\emptyset_{(4,1,1)}$	-0,125	0,023	-5,4	0,0001	$CO_{(t-4)}$
	$\emptyset_{(4,1,2)}$	-0,106	0,035	-2,97	0,003	$PM_{10(t-4)}$
	$\emptyset_{(48,1,1)}$	-0,563	0,025	-22,54	0,0001	$CO_{(t-48)}$
	$\emptyset_{(49,1,1)}$	-0,123	0,025	-4,93	0,0001	$CO_{(t-49)}$
	$\emptyset_{(54,1,1)}$	0,043	0,020	2,12	0,0339	$CO_{(t-54)}$
	$\emptyset_{(96,1,1)}$	-0,306	0,022	-13,56	0,0001	$CO_{(t-96)}$
PM_{10}	$\emptyset_{(1,2,2)}$	0,092	0,027	3,37	0,0008	$PM_{10(t-1)}$
	$\emptyset_{(2,2,1)}$	0,036	0,011	3,09	0,002	$CO_{(t-2)}$
	$\emptyset_{(2,2,2)}$	0,252	0,026	9,38	0,0001	$PM_{10(t-2)}$
	$\emptyset_{(3,2,1)}$	0,045	0,013	3,36	0,0008	$CO_{(t-3)}$
	$\emptyset_{(3,2,2)}$	-0,047	0,020	-2,27	0,0236	$PM_{10(t-3)}$
	$\emptyset_{(4,2,1)}$	0,037	0,013	2,83	0,0047	$CO_{(t-4)}$
	$\emptyset_{(4,2,3)}$	-0,005	0,002	-1,97	0,0496	$NO_{2(t-4)}$
	$\emptyset_{(5,2,1)}$	0,026	0,012	2,22	0,0268	$CO_{(t-5)}$
	$\emptyset_{(48,2,2)}$	-0,683	0,023	-29,33	0,0001	$PM_{10(t-48)}$
	$\emptyset_{(49,2,2)}$	0,082	0,028	2,87	0,0041	$PM_{10(t-49)}$
	$\emptyset_{(50,2,2)}$	0,186	0,029	6,28	0,0001	$PM_{10(t-50)}$
	$\emptyset_{(51,2,1)}$	0,029	0,012	2,44	0,0147	$CO_{(t-51)}$
	$\emptyset_{(54,2,1)}$	-0,035	0,012	-2,84	0,0046	$CO_{(t-54)}$
	$\emptyset_{(96,2,2)}$	-0,511	0,023	-22,15	0,0001	$PM_{10(t-96)}$
	$\emptyset_{(97,2,2)}$	0,141	0,026	5,37	0,0001	$PM_{10(t-97)}$
	$\emptyset_{(98,2,2)}$	0,068	0,027	2,52	0,0117	$PM_{10(t-98)}$
	$\emptyset_{(102,2,1}$	-0,027	0,012	-2,25	0,0249	$CO_{(t-102)}$
NO_2	$\emptyset_{(1,3,1)}$	0,348	0,103	3,37	0,0008	$CO_{(t-1)}$
	$\emptyset_{(1,3,3)}$	-0,459	0,026	-17,33	0,0001	$NO_{2(t-1)}$
	$\emptyset_{(2,3,3)}$	-0,178	0,023	-7,46	0,0001	$NO_{2(t-2)}$

Table 7. Results of VAR Model Parameter Estimation $(5,1,0)(1,1,0)^{48}$ (Continued)

Equation	Parameter	Estimasi	Std Error	t-Value	P-Value	Variabel
NO_2	$\emptyset_{(3,3,3)}$	-0,157	0,024	-6,54	0,0001	$NO_{2(t-3)}$
	$\emptyset_{(4,3,1)}$	-0,231	0,091	-2,51	0,0121	$CO_{(t-4)}$
	$\emptyset_{(4,3,3)}$	-0,078	0,024	-3,16	0,0016	$NO_{2(t-4)}$
	$\emptyset_{(5,3,3)}$	-0,062	0,021	-2,85	0,0045	$NO_{2(t-5)}$
	$\emptyset_{(48,3,1)}$	0,282	0,097	2,88	0,004	$CO_{(t-48)}$
	$\emptyset_{(48,3,3)}$	-0,630	0,025	-24,55	0,0001	$NO_{2(t-48)}$
	$\emptyset_{(49,3,1)}$	0,261	0,114	2,29	0,0225	$CO_{(t-49)}$
	$\emptyset_{(49,3,3)}$	-0,249	0,030	-8,3	0,0001	$NO_{2(t-49)}$
	$\emptyset_{(50,3,1)}$	-0,223	0,088	-2,53	0,0116	$CO_{(t-50)}$
	$\emptyset_{(54,3,3)}$	0,076	0,020	3,76	0,0002	$NO_{2(t-54)}$
	$\emptyset_{(96,3,3)}$	-0,310	0,024	-12,76	0,0001	$NO_{2(t-96)}$
	$\emptyset_{(97,3,3)}$	-0,082	0,025	-3,25	0,0012	$NO_{2(t-97)}$

Based on Table 7, only 41 of 171 significant parameters of the model VAR $(5,1,0)(1,1,0)^{48}$. Furthermore, the VAR $(5,1,0)(1,1,0)^{48}$ model can be written into the following equation.

$$Z_{1,t} = Z_{1,t-1} - 0{,}351Z_{1,t-1} - 0{,}146Z_{1,t-2} - 0{,}159Z_{1,t-3} - 0{,}125Z_{1,t-4} - 0{,}106Z_{2,t-4} - 0{,}563Z_{1,t-48} - 0{,}123Z_{1,t-49} + 0{,}043Z_{1,t-54} - 0{,}306Z_{1,t-96} + a_{1,t}.$$

$$Z_{2,t} = Z_{2,t-1} + 0{,}092Z_{1,t-2} + 0{,}036Z_{1,t-2} + 0{,}252Z_{2,t-2} + 0{,}045Z_{1,t-3} - 0{,}047Z_{2,t-3} + 0{,}037Z_{1,t-4} - 0{,}005Z_{3,t-4} + 0{,}026Z_{1,t-5} - 0{,}683Z_{2,t-48} + 0{,}082Z_{2,t-49} + 0{,}186Z_{2,t-50} + 0{,}029Z_{1,t-51} - 0{,}035Z_{1,t-54} - 0{,}511Z_{2,t-96} + 0{,}141Z_{2,t-97} + 0{,}068Z_{2,t-98} - 0{,}027Z_{1,t-102} + a_{2,t}.$$

$$Z_{3,t} = Z_{3,t-1} + 0{,}348Z_{1,t-1} - 0{,}459Z_{3,t-1} - 0{,}178Z_{3,t-2} - 0{,}157Z_{3,t-3} - 0{,}231Z_{1,t-4} - 0{,}078Z_{3,t-4} - 0{,}062Z_{3,t-5} + 0{,}282Z_{1,t-48} - 0{,}630Z_{3,t-48} + 0{,}261Z_{1,t-49} - 0{,}249Z_{3,t-49} - 0{,}223Z_{1,t-50} + 0{,}076Z_{3,t-54} - 0{,}310Z_{3,t-96} - 0{,}082Z_{3,t-97} + a_{3,t}.$$

Based on the equation above, it can be seen that the results of forecasting the number of indicator levels in a variable are not only influenced by the variable itself but are also influenced by other variables.

4. *Diagnostic Examination of VAR Model*

In the VAR model that is formed with significant parameters, it is necessary to test the assumption of residual *white noise* and the assumption of a multivariate normal distribution of residual assumptions. The results of the diagnostic examination in this study are as follows.

Table 8. Portmanteau Test Results

Lag	Chi-Square	P-value	Lag	Chi-Square	P-value
1	3,1989	0,96203	53	658,4496	0,00099
2	9,9358	0,93706	54	664,3346	0,00099
3	12,9177	0,98901	96	1076,171	0,00099
4	18,1319	0,99400	97	1083,4327	0,00099
5	24,7305	0,99500	98	1090,0658	0,00099
48	640,0940	0,00099	99	1100,9041	0,00099
49	642,8955	0,00099	100	1108,1198	0,00099
50	645,6256	0,00099	101	1113,1186	0,00099
51	649,2034	0,00099	102	1121,5494	0,00099
52	654,2498	0,00099			

Based on Table 8, the residuals from the data do not meet the assumption of residual *white noise*. This can be seen through the *p-value* in some *lags*, which are less than 0,05. Next is the examination of the distribution of residuals *multivariate* normal.

Table 9. Shapiro-Wilk Test Results

W	*P-Value*
0, 88339	$< 2{,}2 \times 10^{-16}$

Test results using Shapiro-Wilk produce *p-value* a significantly smaller than 0,05, so it can be concluded that the residuals from the data do not meet the assumption that the residuals have a multivariate normal distribution. The VAR $(5,1,0)(1,1,0)^{48}$ model does not meet the assumption of residual *white noise* and has a multivariate normal distribution, so it cannot be used to predict and determine the relationship between air quality determinants at the SUF 7 station in Surabaya. Based on this, the detection will be carried out *for the outlier*.

D. Outlier Detection

Outlier detection can be performed using residual control charts are calculated by statistical T^2-*Hotelling* and performed *iteratively* until the condition is found *in control*. The results of outlier detection on indicator data for determining air quality at SUF 7 station in Surabaya can be seen as follows.

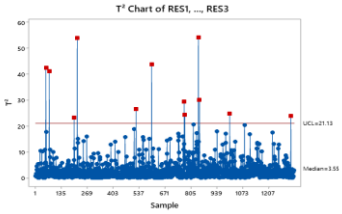

Figure 7. Residual Control Diagram Before Iteration

In the residual control diagram before iteration, it was detected that there were 12 signals out of control, namely the 55, 73,201,217,521, 602, 770, 773, 843, 849, 1005, and 1321 residual data and then iterated by removing outlier data residuals as can be seen in Figure 8.

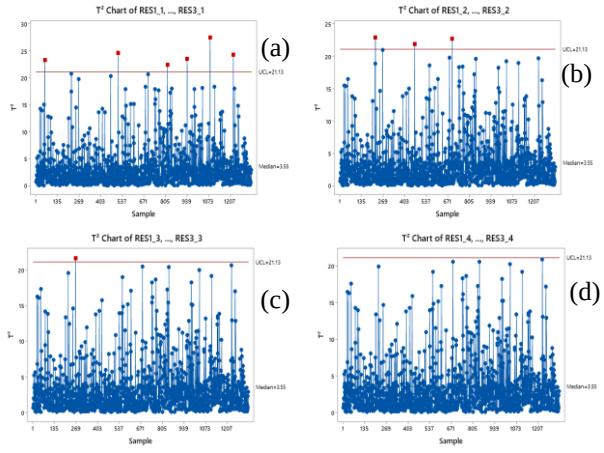

Figure 8. (a) Residual Control Diagram After the First Iteration (b) Second Iteration (c) Third Iteration and (d) Fourth Iteration

After the first iteration to the fourth iteration, the residual control diagram detects a total of 22 out of control signals, namely the residual data 57, 511, 817, 940, 1082, 12 26, 221, 464, 697, and 266. The results of the residual control diagram after the fourth iteration can be seen in Figure 8 part (d). Because the residuals are already in control, then the dummy variable is formed on the data used.

Table 10. Variables Dummy Model VAR (5,1,0)(1,1,0)[48]

Residual	D_1	D_2	D_3	⋯	D_{22}
55	1	0	0	⋯	0
73	0	1	0	⋯	0
201	0	0	1	⋯	0
⋮	⋮	⋮	⋮	⋱	⋮
266	0	0	0	⋯	1

After the dummy variables have been created, then a new model is created to obtain the appropriate model using VARX.

E. Modeling Determinant Indicators of Air Quality at SSUF 7 Station Surabaya with (VARX)

The stages carried out in modeling using VARX can be seen in the following sub-sections.

1. *Identification of VARX Model*

Identification of the VARX model can be seen through *Matrix Partial Cross-Correlation Function* (MPCCF) plots. The MPCCF plot in this study with variables *dummy* can be seen in Figure 9.

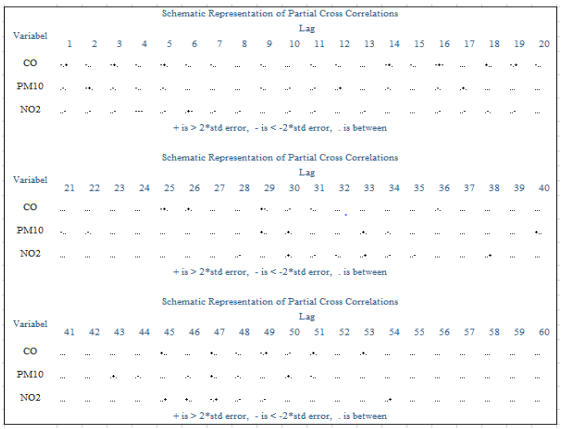

Figure 9. MPCCF Plot of Three Variables and variables d*ummy*

Based on Figure 9, the determinants of air quality at SUF 7 Station Surabaya are significant in the 1st lag to the 5th lag. So the VARX model formed is (5,5).

2. *Parameter Estimation and Significance Test Parameter*

Estimation of the VARX model was carried out on the VAR model (5,5) and found as many as 297 parameters. Furthermore, a significance test was conducted using *restrict* to determine the significant parameters to the model.

Table 11. Result of Parameter Estimation of VARX Model (5,5)

Equation	Parameter	Estimasi	Std Error	t-Value	P-Value	Variabel
CO	$\delta_{(0,1,7)}$	-0,584	0,109	-5,34	0,0001	$D_{7(t)}$
	$\delta_{(1,1,3)}$	-0,220	0,109	-2,02	0,0436	$D_{3(t-1)}$
	$\delta_{(2,1,3)}$	-0,290	0,117	-2,47	0,0137	$D_{3(t-2)}$
	$\delta_{(3,1,7)}$	-0,558	0,109	-5,09	0,0001	$D_{7(t-3)}$
	$\delta_{(4,1,6)}$	0,255	0,109	2,35	0,0191	$D_{6(t-4)}$

Tabel 11. Result of Parameter Estimation of VARX Model (5,5) (Continued)

Equation	Parameter	Estimasi	Std Error	t-Value	P-Value	Variabel
CO	$\delta_{(4,1,7)}$	0,324	0,110	2,95	0,0033	$D_{7(t-4)}$
	$\delta_{(5,1,3)}$	-0,222	0,108	-2,04	0,0413	$D_{3(t-5)}$
	$\delta_{(5,1,7)}$	-0,372	0,110	-3,38	0,0007	$D_{7(t-5)}$
	$\emptyset_{(1,1,1)}$	0,641	0,026	24,39	0,0001	$CO_{(t-1)}$
	$\emptyset_{(2,1,1)}$	0,191	0,029	6,47	0,0001	$CO_{(t-2)}$
	$\emptyset_{(3,1,1)}$	-0,102	0,025	-4,02	0,0001	$CO_{(t-3)}$
	$\emptyset_{(48,1,1)}$	0,179	0,019	9,26	0,0001	$CO_{(t-48)}$
	$\emptyset_{(52,1,2)}$	0,108	0,039	2,75	0,0061	$PM_{10(t-52)}$
	$\emptyset_{(53,1,1)}$	-0,087	0,020	-4,27	0,0001	$CO_{(t-53)}$
	$\emptyset_{(53,1,2)}$	-0,097	0,037	-2,59	0,0098	$PM_{10(t-53)}$
	$\emptyset_{(54,1,3)}$	0,012	0,003	4,02	0,0001	$NO_{2(t-54)}$
	$\emptyset_{(96,1,1)}$	0,074	0,024	3,05	0,0024	$CO_{(t-96)}$
	$\emptyset_{(96,1,2)}$	0,109	0,037	2,89	0,0039	$PM_{10(t-96)}$
	$\emptyset_{(97,1,1)}$	0,064	0,028	2,24	0,0256	$CO_{(t-97)}$
	$\emptyset_{(97,1,2)}$	-0,093	0,038	-2,44	0,0148	$PM_{10(t-97)}$
	$\emptyset_{(98,1,1)}$	-0,116	0,025	-4,52	0,0001	$CO_{(t-98)}$
	$\emptyset_{(102,1,1)}$	-0,045	0,018	-2,47	0,0138	$CO_{(t-102)}$
PM_{10}	$\delta_{(0,2,1)}$	0,437	0,058	7,49	0,0001	$D_{1(t)}$
	$\delta_{(0,2,2)}$	-0,440	0,058	-7,55	0,0001	$D_{2(t)}$
	$\delta_{(0,2,3)}$	0,309	0,058	5,33	0,0001	$D_{3(t)}$
	$\delta_{(0,2,4)}$	-0,487	0,058	-8,38	0,0001	$D_{4(t)}$
	$\delta_{(0,2,5)}$	0,263	0,058	4,53	0,0001	$D_{5(t)}$
	$\delta_{(0,2,6)}$	-0,422	0,058	-7,26	0,0001	$D_{6(t)}$
	$\delta_{(2,2,1)}$	0,297	0,059	5,03	0,0001	$D_{1(t-2)}$
	$\delta_{(2,2,2)}$	-0,132	0,059	-2,22	0,0269	$D_{2(t-2)}$
	$\delta_{(2,2,4)}$	-0,259	0,059	-4,36	0,0001	$D_{4(t-2)}$
	$\delta_{(4,2,2)}$	0,152	0,059	2,58	0,0101	$D_{2(t-4)}$
	$\delta_{(4,2,4)}$	0,275	0,059	4,61	0,0001	$D_{4(t-4)}$
	$\delta_{(4,2,6)}$	0,127	0,059	2,15	0,0316	$D_{6(t-4)}$
	$\emptyset_{(1,2,1)}$	0,064	0,008	7,29	0,0001	$CO_{(t-1)}$
	$\emptyset_{(1,2,2)}$	1,093	0,024	45,42	0,0001	$PM_{10(t-1)}$
	$\emptyset_{(2,2,2)}$	0,166	0,036	4,55	0,0001	$PM_{10(t-2)}$
	$\emptyset_{(3,2,2)}$	-0,388	0,034	-11,12	0,0001	$PM_{10(t-3)}$
	$\emptyset_{(4,2,2)}$	0,161	0,036	4,47	0,0001	$PM_{10(t-4)}$
	$\emptyset_{(5,2,2)}$	-0,071	0,023	-2,98	0,0029	$PM_{10(t-5)}$
	$\emptyset_{(48,2,2)}$	0,067	0,014	4,65	0,0001	$PM_{10(t-48)}$
	$\emptyset_{(50,2,2)}$	-0,065	0,014	-4,57	0,0001	$PM_{10(t-50)}$
	$\emptyset_{(52,2,1)}$	-0,023	0,009	-2,37	0,0182	$CO_{(t-52)}$
	$\emptyset_{(54,2,3)}$	0,004	0,001	3,06	0,0023	$NO_{2(t-54)}$
	$\emptyset_{(96,2,1)}$	0,034	0,008	3,8	0,0002	$CO_{(t-96)}$
	$\emptyset_{(97,2,2)}$	0,173	0,021	8,03	0,0001	$PM_{10(t-97)}$
	$\emptyset_{(98,2,2)}$	-0,171	0,021	-7,96	0,0001	$PM_{10(t-98)}$
	$\emptyset_{(99,2,1)}$	-0,039	0,010	-3,67	0,0003	$CO_{(t-99)}$
	$\emptyset_{(102,2,1)}$	0,048	0,009	5,01	0,0001	$CO_{(t-102)}$
NO2	$\delta_{(0,3,5)}$	-1,527	0,451	-3,38	0,0007	$D_{5(t)}$
	$\delta_{(2,3,2)}$	1,152	0,452	2,55	0,011	$D_{2(t-2)}$
	$\delta_{(2,3,3)}$	-1,044	0,487	-2,14	0,0323	$D_{3(t-2)}$
	$\delta_{(3,3,1)}$	0,897	0,452	1,98	0,0476	$D_{1(t-3)}$
	$\delta_{(3,3,2)}$	0,936	0,452	2,07	0,0387	$D_{2(t-3)}$
	$\emptyset_{(1,3,1)}$	0,458	0,086	5,27	0,0001	$CO_{(t-1)}$

Tabel 11. Result of Parameter Estimation of VARX Model (5,5) (Continued)

Equation	Parameter	Estimasi	Std Error	t-Value	P-Value	Variabel
	$\emptyset_{(1,3,3)}$	0,477	0,026	18,22	0,0001	$NO_{2(t-1)}$
	$\emptyset_{(2,3,3)}$	0,199	0,026	7,4	0,0001	$NO_{2(t-2)}$
	$\emptyset_{(4,3,1)}$	-0,344	0,089	-3,87	0,0001	$CO_{(t-4)}$
	$\emptyset_{(4,3,3)}$	0,068	0,022	2,99	0,0028	$NO_{2(t-4)}$
	$\emptyset_{(48,3,3)}$	0,136	0,020	6,65	0,0001	$NO_{2(t-48)}$
	$\emptyset_{(49,3,2)}$	0,361	0,170	2,12	0,0341	$PM_{10(t-49)}$
	$\emptyset_{(50,3,2)}$	-0,490	0,198	-2,48	0,0133	$PM_{10(t-50)}$
	$\emptyset_{(51,3,1)}$	0,288	0,100	2,87	0,0042	$CO_{(t-51)}$
	$\emptyset_{(51,3,3)}$	-0,056	0,023	-2,44	0,0146	$NO_{2(t-51)}$
	$\emptyset_{(53,3,1)}$	-0,213	0,096	-2,22	0,0264	$CO_{(t-53)}$
	$\emptyset_{(54,3,2)}$	0,217	0,069	3,14	0,0017	$PM_{10(t-54)}$
	$\emptyset_{(54,3,3)}$	0,065	0,019	3,3	0,001	$NO_{2(t-54)}$
	$\emptyset_{(96,3,3)}$	0,098	0,022	4,47	0,0001	$NO_{2(t-96)}$
	$\emptyset_{(98,3,1)}$	0,306	0,094	3,23	0,0013	$CO_{(t-98)}$
	$\emptyset_{(98,3,3)}$	-0,062	0,023	-2,7	0,007	$NO_{2(t-98)}$
	$\emptyset_{(99,3,2)}$	-0,385	0,163	-2,36	0,0184	$PM_{10(t-99)}$
	$\emptyset_{(100,3,1)}$	-0,197	0,089	-2,2	0,028	$CO_{(t-100)}$
	$\emptyset_{(100,3,3)}$	0,377	0,161	2,34	0,0195	$PM_{10(t-100)}$

Based on Table 11, it can be seen that there are only 73 of 297 parameters that are significant to the VARX model (5,5) after *restricting* 224 parameters. Furthermore, the VARX model (5,5) can be written into the following equation.

$$Z_{1,t} = -0{,}584D_{7,t} - 0{,}220D_{3,t-1} - 0{,}290D_{3,t-2} - 0{,}558D_{7,t-3} + 0{,}255D_{6,t-4} + 0{,}324D_{7,t-4} - 0{,}222D_{3,t-5} - 0{,}372D_{7,t-5} + Z_{1,t-1} + 0{,}641Z_{1,t-1} + 0{,}191Z_{1,t-2} - 0{,}102Z_{1,t-3} + 0{,}179Z_{1,t-48} + 0{,}108Z_{2,t-52} - 0{,}087Z_{1,t-53} - 0{,}097Z_{2,t-53} + 0{,}012Z_{3,t-54} + 0{,}074Z_{1,t-96} + 0{,}109Z_{2,t-96} + 0{,}064Z_{1,t-97} - 0{,}093Z_{2,t-97} - 0{,}116Z_{1,t-98} - 0{,}045Z_{1,t-102} + a_{1,t}.$$

$$Z_{2,t} = 0{,}437D_{1,t} - 0{,}440D_{2,t} + 0{,}309D_{3,t} - 0{,}487D_{4,t} + 0{,}263D_{5,t} - 0{,}422D_{6,t} + 0{,}297D_{1,t-2} - 0{,}132D_{2,t-2} - 0{,}259D_{4,t-2} + 0{,}152D_{2,t-4} + 0{,}275D_{4,t-4} + 0{,}127D_{6,t-4} + Z_{2,t-1} + 0{,}064Z_{1,t-1} + 1{,}093Z_{2,t-1} + 0{,}166Z_{2,t-2} - 0{,}388Z_{2,t-3} + 0{,}161Z_{2,t-4} - 0{,}071Z_{2,t-5} + 0{,}067Z_{2,t-48} - 0{,}065Z_{2,t-50} - 0{,}023Z_{1,t-52} + 0{,}004Z_{3,t-54} + 0{,}034Z_{1,t-96} + 0{,}173Z_{2,t-97} - 0{,}171Z_{2,t-98} - 0{,}039Z_{1,t-99} + 0{,}048Z_{1,t-102} + a_{2,t}.$$

$$Z_{3,t} = -1{,}527D_{5,t} + 1{,}152D_{2,t-2} - 1{,}044D_{3,t-2} + 0{,}897D_{1,t-3} + 0{,}936D_{2,t-3} + Z_{3,t-1} + 0{,}458Z_{1,t-1} + 0{,}477Z_{3,t-1} + 0{,}199Z_{3,t-2} - 0{,}344Z_{1,t-4} + 0{,}068Z_{3,t-4} + 0{,}136Z_{3,t-48} + 0{,}361Z_{2,t-49} - 0{,}490Z_{2,t-50} + 0{,}288Z_{1,t-51} - 0{,}056Z_{3,t-51} - 0{,}213Z_{1,t-53} + 0{,}217Z_{2,t-54} + 0{,}065Z_{3,t-54} + 0{,}098Z_{2,t-96} + 0{,}306Z_{1,t-98} - 0{,}062Z_{3,t-98} - 0{,}385Z_{2,t-99} - 0{,}197Z_{1,t-100} + 0{,}377Z_{2,t-100} + a_{3,t}.$$

Based on the above equation, it can be seen that the results of forecasting the number of indicator levels in a variable are not only influenced by the variable itself but are also influenced by other variables and variables *dummy*.

3. *Diagnostic Examination of the VARX Model*

In the VARX model that was formed, the next step is testing the assumption of residual *white noise* and the multivariate normal distribution. The results of the diagnostic examination in this study are as follows.

Table 12. Portmanteau Test Results VARX Model (5,5)

Lag	Chi-Square	P-value	Lag	Chi-Square	P-value
1	6,2048	0,73926	53	706,1395	0,00099
2	17,4680	0,52247	54	713,4031	0,00099
3	22,4179	0,69430	96	1143,7571	0,00099
4	26,8656	0,85514	97	1148,6326	0,00099
5	31,8352	0,92307	98	1158,8571	0,00099
48	679,9916	0,00099	99	1170,2791	0,00099
49	685,3938	0,00099	100	1182,6872	0,00099
50	688,6889	0,00099	101	1187,9186	0,00099
51	693,8673	0,00099	102	1197,1581	0,00099
52	698,4971	0,00099			

Based on Table 12, the residuals from the data used to meet the assumption of residual *white noise* in five *the lags* initial used and in *the lags* subsequent do not meet the assumption of residual *white noise*. This can be seen through the *p-value*, which is less and more than 0,05, so that the residual data used in this study does not meet the assumption of residual *white noise* but is assumed to have met the assumption of residual *white noise*. Next is the examination of the distribution of residuals *multivariate* normal.

Table 13. Shapiro-Wilk Test Results

W	*P-Value*
0,90967	$< 2{,}2 \times 10^{-16}$

Test results using Shapiro-Wilk produce a very small *p-value*, less than 0,05, so it can be concluded that the residuals from the data on air quality determinants at SUF 7 station Surabaya do not meet the residual assumptions normal multivariate distribution. However, in this study, the residuals are assumed to have met the assumption of normal multivariate distribution residuals and can be used for the next analysis stage.

F. Forecasting Results and Selection of The Best Model

After modeling, the next thing that needs to be done is forecasting. The plot of forecasting results for the indicators of CO, PM_{10}, and NO_2 on the VAR and VARX methods can be seen as follows.

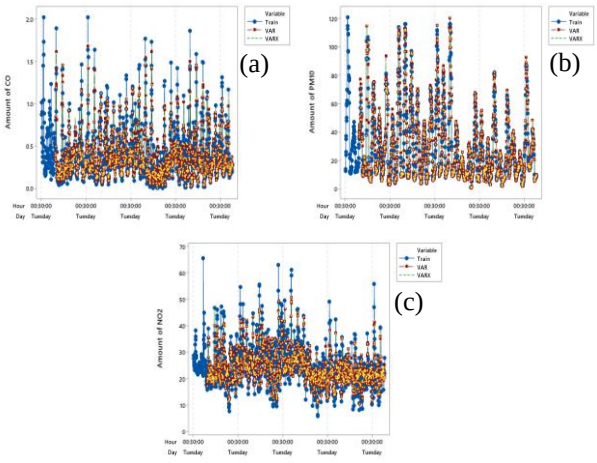

Figure 10. Results of Forecasting Data *Train* Indicator (a) CO (b) PM_{10} (c) NO_2

The results of forecasting data *train* using the VAR and VARX methods on the three air quality determinants at the SUF 7 station in Surabaya have forecasting results that follow actual data. Furthermore, the results of the data forecasting will be seen *test* on the three indicators that determine air quality at the SUF 7 station in Surabaya.

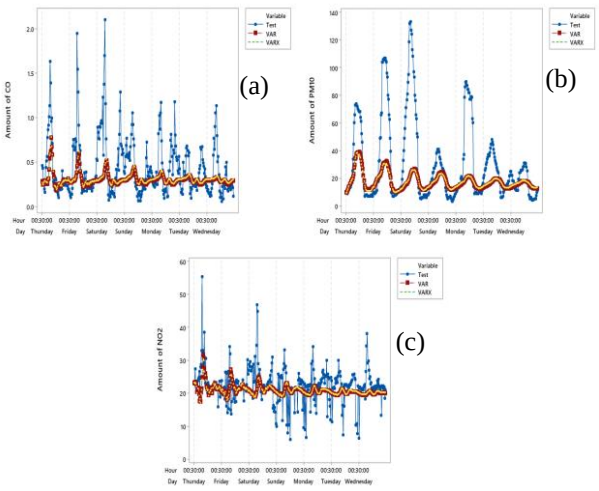

Figure 11. Results of Forecasting Data *Test* Indicator (a) CO (b) PM_{10} (c) NO_2

The results of the forecasting data *test* using the VAR and VARX methods on the CO, PM_{10}, and NO_2 indicators tend to follow the actual data. After knowing the results of forecasting on each indicator determining air quality at SUF 7 Station in Surabaya, then the best model will be selected. The selection of the best model is made by looking at the smallest RMSE value generated by each model. The RMSE values for each indicator and method used in this study can be seen in Table 14.

Table 14. The Smallest RMSE Value for Each Method

Data	Indicator	Method	RMSE
Train	CO	VAR	0,147470914
		VARX	**0,14387047**
	PM_{10}	VAR	3,150165032
		VARX	**2,6465331**
	NO_2	VAR	4,597958256
		VARX	**4,54182874**
Test	CO	VAR	0,322961028
		VARX	**0,31755301**
	PM_{10}	VAR	**29,35093274**
		VARX	29,40259195
	NO_2	VAR	**5,72878423**
		VARX	5,75267423

The best forecasting results for data *train* on the CO, PM_{10}, and NO_2 indicators are forecasting results derived from the method VARX, and this can be known based on the value of the goodness of the model or the resulting RMSE value, which has the smallest value compared to other methods used as can be seen in Table 14. Forecasting results on the data *test* show that the VARX method produces the best forecasts for the CO indicator and the VAR method produces the best forecast for PM_{10} and NO_2 indicators.

V. CONCLUSIONS AND SUGGESTIONS

A. *Conclusions*

Based on the results of the analysis and discussion that have been described, the conclusions obtained in this study are as follows.

1. The three indicators that determine air quality, namely CO, PM_{10}, and NO_2 at SUF 7 station in Surabaya, have a *seasonal* pattern, namely daily seasonality per half hour. The total levels of the three indicators tend to be low in the afternoon until the evening and high in the morning until the afternoon.
2. In the VAR method, the VAR $(5,1,0)(1,1,0)^{48}$ model is obtained. The residuals of the $(5,1,0)(1,1,0)^{48}$ model do not meet the assumption of residual *white noise* and have distribution *multivariate* normal, so it is necessary to detect *outliers* in the data used.
3. Outlier detection is done using the residual control charts procedure *iterative* with statistical T^2 *Hoteling* and found to be 22 *outliers* iterations four times. Furthermore, variables were created *dummy* as many as *outliers* in the data and modeled using VARX.
4. The VARX method produces a VARX model (5,5). The residuals from the data in VARX do not meet the assumption of residual *white noise* and have distributions *multivariate* normal but are assumed to have met the assumptions of normal and multivariate distributions of residuals *white noise*.
5. The best model obtained in this study came from the results of modeling using VARX. This can be seen through the results of forecasting from the VARX method on the data *train* in this study, which produced the smallest value for the goodness of the model for the three indicators that determine air quality compared to the forecasting results from the VAR method.

B. *Suggestions*

Based on the research results that have been obtained, suggestions from the authors that can be given as consideration for further research are as follows.

1. There are many missing data, so it is necessary to improve the data collection management in order to produce better models and forecasts.
2. Perform detection *outlier* using *Multivariate Cumulative Sum* (MCUSUM) and *Multivariate Exponentially Weighted Moving Average* (MEWMA).
3. It is necessary to have policies made by the Surabaya Government to prevent air pollution in Surabaya.

REFERENCES

[1] Pemerintah Kota Surabaya, "Transportasi dan Pariwisata," in *Statistik Sektoral Kota Surabaya Tahun 2019*, Surabaya, 2019, p. 432.

[2] S. Fardiaz, Polusi Air dan Udara, Yogyakarta: Kanisius, 2003.

[3] D. Muziansyah, S. Rahayu and S. Sebayang, "Model Emisi Gas Buangan Kendaraan Bermotor Akibat

Aktivitas Transportasi (Studi Kasus: Terminal Pasar Bawah Ramayana Kota Bandar Lampung)," *Jurnal Rekayasa Sipil dan Desain*, pp. 57-70, 2015.

[4] D. N. Wijayanti, Gambaran dan analisis Risiko Nitrogen Dioksida (NO2) per Kota/Kabupaten dan Provindi di Indonesia (Hasil Pemantauan Kualitas Udara Ambien dengan Metode Pasif di Pusarpedal Tahun 2011), Jakarta: Skripsi Universitas Indonesia, 2012.

[5] IQ Air, "IQ Air," 2021. [Online]. Available: https://www.iqair.com/id/indonesia/east-java/surabaya. [Accessed 23 February 2021].

[6] R. A. Robles, J. C. Ortega, J. S. Fu and G. Reed, "A Hybrid ARIMA and Artificial Neural Network Model to Forecast Particullate Matter in Urban Areas: The Case of Temuco, Chile," *Journal Atmosphere Environment,* no. 42, pp. 8331-8440, 2008.

[7] H. Prabowo and Suhartono, Peramalan Kualitas Udara di Kota Surabaya untuk menentukan Kategori Indeks Standar Pencemar Udara, Surabaya: Tugas Akhir Departemen Statistika ITS, 2019, pp. 1-5.

[8] B. Chrisdayanti, and A. Suharsono, "Peramalan Kandungan Partikular Matter (PM10) dalam Udara Ambien Kota Surabaya Menggunakan Double Seasonal ARIMA," *Jurnal Sains dan Seni ITS,* no. 4(2), pp. 242-247, 2015.

[9] N. Nahdliyah, Suhartono and M. S. Akbar, Model Multivariat Generalized Space Time Autoregressive (MGSTAR) untuk Monitoring Kualitas Udara di Surabaya, Surabaya: Tugas Akhir Departemen Statistika ITS, 2019, pp. 79-104.

[10] G. E. Box, G. M. Jenkins and G. C. Reinsel, Time Series Analysis: Forecasting and Control Third Edition, vol. 62, Englewood Cliffs: Prentice Hall, 1994, pp. 540-552.

[11] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via The Expectation Maximization (EM) Algorithm," *Journal of The Royal Statistical Society Series B (Methodological),* pp. 1-38, 1977.

[12] D. Nettleton, "Selection of Variables and Factor Derivation," in *Commercial Data Mining*, 2014, pp. 79-104.

[13] W. S. Wei, Time Series Analysis: Univariate and Multivariate Methods, New York: Pearson Education, 2006.

[14] R. S. Tsay, Multivariate Time Series Analysis, Chicago: John WIley, Inc, 2014.

[15] H. Lutkepohl, New Introduction to Multiple Time Series Analysis, New York: Springer, 2005, pp. 591-611.

[16] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Sample)," *Biometriks,* pp. 591-611, 1965.

[17] R. Oduk, Control Chart for Serially Dependent Multivariate Data, vol. 1, Technical University Denmark, 2012, pp. 2535-2537.

[18] A. Suharsono, S. Guritno and Subanar, "Autoregressive Vector Modelling Simulation with Innovative Outlier," *Journal of Basic and Applied Scientific Research,* vol. 1, no. 12, pp. 2535-2537, 2011.

[19] A. Suharsono, Pemodelan Vector Autoregressive dengan Adanya Outlier atau Pergeseran terhadap Rata-rata, Yogyakarta: FMIPA Universitas Gajah Mada, 2012.

[20] Menteri Negara Lingkungan Hidup, Keputusan Menteri Negara Lingkungan Hidup No. 45 Tahun 1977 tentang Indeks Standar Pencemar Udara, Jakarta, 1997.