**Journal of Theoretical and Applied Information Technology**
31st January 2023. Vol.101. No 2
© 2023 Little Lion Scientific

ISSN: **1992-8645** www.jatit.org E-ISSN: **1817-3195**

# AIR QUALITY CONTROL USING $T^2$ HOTELLING DOUBLE BOOTSTRAP CONTROL CHART WITH VECTOR AUTOREGRESSIVE MODEL APPROACH

**JAUHARIN INSIYAH[1] , SUCI ASTUTIK[2], LOEKITO ADI SOEHONO[3]**
[123]Departement of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University, Malang 65145, Indonesia
E-mail: [1]jauharininisyah@gmail.com, [2]suci_ub@ub.ac.id, [3]loekito@loekito.id

## ABSTRACT

This study aims to find a sensitive control chart to shifts in the control process that induces a correlation among time series observations, known as autocorrelation. $T^2$ Hotelling, a popular multivariate chart, is no longer sensitive to detect small and moderate mean shifts derived from the autocorrelation process. Therefore, this study uses Vector Autoregressive Model (VAR) residuals to build $T^2$ Hotelling control charts. To improve the chart Double Bootstrap method is used to construct a sensitive control limit because the assumptions of the $T^2$ Hotelling are not fulfilled. Violation of assumptions results in the analysis being inappropriate. The proposed control chart is used for air quality control in Surabaya with the characteristics quality of PM 2.5, PM 10, and CO, which are correlated with each other. The proposed control chart's performance is compared with the single Bootstrap control chart by Average Run Length (ARL) value at different numbers of observations. The results show that The Proposed $T^2$ based on residual VAR with Double Bootstrap is more sensitive than the single Bootstrap to detect out-of-control on all shifts and at observations. Thus, the proposed control chart can be a way to minimize errors in controlling air quality.

**Keywords:** *Air Quality Control, Double Boostrap, Multivariate Control Chart, $T^2$ Hotelling, Vector Autoregressive (VAR)*

## 1. INTRODUCTION

According to the latest report by the Air Quality Life Index (AQLI), which is a world organization that focuses on air quality, it is stated that Indonesia is ranked 13th as the most polluting country out of 243 countries in the world [1]. Even based on the air quality threshold the World Health Organization (WHO) applied, 9 out of 10 people in Indonesia have been exposed to air pollution. It decreases life expectancy in Indonesia by 1.2 years due to air pollution [2].

Air pollution includes six pollutants, mainly Carbon Monoxide (CO), Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), surface ozone (O3), and Particulate Matter (PM), which include PM 2.5 and PM 10. Among the six, PM 2.5, PM 10, and CO are the primary pollutants that contribute to the main air pollution. The third source is caused by motor vehicle emissions, coal, industrial plants, and biomass burning. Exposure to all three can cause various health problems, such as lung cancer, ischemic heart disease, neurological disorders, stroke, and nerve problems. So serious attention is needed in handling air quality in a region.

Previous studies stated that PM 2.5, PM 10, and CO had a fairly close correlation [3] [4]. However, the air quality analysis is done individually. When analyzing the concentration separately, even though there is a statistical correlation between the two, one can get a less accurate prediction [5]. Therefore, in this study, air quality analysis was carried out using a Statistical Process Control approach. A popular method used in process control is the $T^2$ Hotelling control chart. This control chart can control two or more characteristics simultaneously. The $T^2$ Hotelling control chart has limitations; it can only be applied to multivariate normally distributed data, while pollutants PM 2.5, PM 10, and CO are time-series data containing autocorrelation. In reference [6], autocorrelation strongly and negatively impacts the $T^2$ Hotelling control chart. The approach that can be taken if there is autocorrelation in the data is to use residuals from the time series model, which is Vector Autoregressive (VAR). The residuals generated by the best time series model will meet the independent and identical assumptions [7]. The residuals obtained from the best VAR model were then analyzed using the $T^2$ Hotelling control chart.

The residual VAR diagram research conducted by [8][9] shows that apart from being effective in monitoring multivariate processes that contain autocorrelation, it is also sensitive to small changes in a single parameter effect on the entire system. Reference [10] prove that the $T^2$ Hotelling chart based on residuals from the VAR model detects shift faster on the means process and is effective for negative autocorrelation with larger shifts than a standard deviation.

Control limits are the main thing in the quality control process using control charts. In data with autocorrelation, not all control charts can detect shifts correctly. One of the control limits that was developed by previous research is the Bootstrap method approach. Bootstrapping is a method that takes and returns from samples (resampling) observations representing the original population. In the Bootstrap Control Chart, the control limit is calculated based on the confidence interval of the $T^2$ statistic that has been bootstrapped. Bootstrap was first introduced by [11] with one advantage does not have to follow a normal distribution. Bjeger first used the bootstrap control chart on the Shewhart control chart [12]. The other researchers also developed a bootstrap method for control charts that follow the Weibull, Birnbaum Sanders, and Inverse Gaussian distributions [13] [14] [15].

Phaladiganon *et al.* [9] used the bootstrap method for the Multivariate Hotelling control chart (in this study called $T^2$ PB Chart). By comparing ARL values, Phaladiganon *et al.* show that compared to Kernel Density Estimation (KDE), the bootstrap method is more effective in normal and non-normal situations. Reference [16] shows that the bootstrap control limit on the $T^2$ Hotelling control chart performs better than KDE. Mostajeran *et al.* [17] compared the bootstrap control chart with the Sign and Wilcoxon control chart. Using a bootstrap control chart on a lognormal distribution can also detect out of control quickly[18].

Then Mostajeran *et al.* [19] proposed a $T^2$ Hotelling chart with a different bootstrap method from that proposed by Phaladiganon. Mostajeran *et al.* proposed that the New Bootstrap control chart (in this study called $T^2$ NB Chart) uses [11] bootstrap principle, which resamples the original data. $T^2$ NB Chart first resamples the original data and then calculates the statistics on the resampling data to compute the control limit. This comparison shows that the $T^2$ NB Chart performs better on small data than the $T^2$ PB Chart on large data. However, they both perform better than the classic $T^2$ Hotelling. Therefore, this study proposes a new method for the $T^2$ Hotelling control chart with the Double Bootstrap approach. Double Bootstrap was introduced by Beran [20] with better performance than single Bootstrap. The principle is that the first stage Bootstrap dataset is replicated as much as $B_1$ the original dataset, and the bootstrapping process is carried back as much as $B_2$ replication. In reference [21], Double Bootstrap can detect out-of-control with small ARL values.

A hybrid method of the Double Bootstrap approach is proposed in this study to obtain more sensitive results using a $T^2$ control chart based on residual VAR. The residuals generated from the VAR model in this study were analyzed with an $T^2$ NB Chart control chart in the first step, which is called a single Bootstrap ($T^2_{VAR}$ NB Chart), and a $T^2$ PB Chart ( $T^2_{VAR}$ PB Chart) was used to build a control limit then called double Bootstrap ($T^2_{VAR}$ DB Chart). The proposed study aims to increase the sensitivity of the $T^2$ Hotelling control chart in detecting out-of-control points on air quality data in Surabaya. In addition, the proposed study is also effectively used in various numbers of observations. The performance of the proposed study is seen from the estimated value of the Average Run length (ARL) using the selection of shifts in the $T^2_{VAR}$ NB Chart, $T^2_{VAR}$ PB Chart and $T^2_{VAR}$ DB Chart.

In this paper, Section 2 describes the literature of $T^2$ control charts based on residual model VAR. Also, describe the control limit of $T^2$ Hotelling control chart using the single and double bootstrap approach. Section 3 presents the dataset and displays the performance of the comparison of the proposed chart. Finally, section 4 shows the summarized results.

## 2. LITERATURE REVIEW

Vector Autoregressive (VAR) is a method for analyzing time series data resulting from the development of Autoregressive (AR). The VAR model can model data with two or more interrelated variables [22]. In control charts, VAR is an approach that has been proven to overcome autocorrelation in data by modeling the data with an appropriate time series model. The model was created to eliminate autocorrelation in the data and apply the residuals to the control chart [23].

## 2.1 Correlation Test

The test statistics for correlation testing are as follows [24].

$$\chi^2_{statistic} = -\left\{n - 1 - \frac{2m+5}{6}\right\}\ln|\mathbf{R}| \qquad (1)$$

Explanation:
$m$ : number of quality characteristics
$n$ : number of samples
$\mathbf{R}$ : correlation matrix of each quality characteristic
Inter-residual quality characteristics are said to be correlated if $\chi^2_{statistic} > \chi^2_{m(m-1)/2}$.

## 2.2 Vector Autoregressive

### 2.2.1 Stationery

In time series modeling, the data must first be stationary concerning the variance and average, that is, when the standard is $E(Z_t) = \mu$ and the variance $Var(Z_t) = E(Z_t) = E(Z_t - \mu)^2 = \sigma^2$ constant [25].

For the variance $T(Z_t)$ constant, a transformation is carried out based on the value that meets equation (5) below.

$$T(\mu_t) = \int \frac{1}{\sqrt{f(\mu_t)}} d\mu_t \qquad (2)$$

Meanwhile, the stationarity test against the average was carried out using the Augmented Dickey-Fuller (ADF) test. The test statistics in the ADF test are as follows.

$$\tau = \frac{\phi - 1}{S_\phi} \qquad (3)$$

Where:
$\phi$ : AR parameter estimate value
$\tau$ : standard error value for the predicted value of the parameter $\phi$
With,

$$\hat{\phi} = \frac{\sum_{t=1}^{n} Z_{t-1} - Z_t}{\sum_{t=1}^{n} Z_{t-1}^2}$$

$$S_{\hat{\phi}} = \left[\overset{\square}{\sigma}{}^2_\alpha \left(\sum_{t-1}^{n} Z_{t-1}^2\right)^{-1}\right]^{1/2}$$

$$\overset{\square}{\sigma}{}^2_\alpha = \sum_{t-1}^{n} \frac{(Z_t - \hat{\phi} Z_{t-1})^2}{(n-1)}$$

If $\tau > \tau_{\alpha,n}$ then accepted $H_0$, time-series data is not stationery concerning the average. If $\tau > \tau_{\alpha,n}$ rejected $H_0$, is the stationary time-series data to the average [22].

### 2.2.2 Model Vector Autoregressive

The general form of the VAR model is as follows

$$Z_t = \mu + \Phi_1 Z_{t-1} + \ldots + \Phi_p Z_{t-p} + a_t$$

The order of the AR vector model can be determined from the partial autocorrelation matrix function. For example, a partial autocorrelation matrix on lag s is obtained from the following equation [25].

$$\mathbf{P}(s) = [\mathbf{D}_v(s)]^{-1} \mathbf{V}_{\mathrm{Vu}}(s) [\mathbf{D}_u(s)]^{-1} \qquad (4)$$

where:

$$\mathbf{V}_u(s) = var(\mathbf{u}_{s-1,t+s}) = \Gamma(0) - \boldsymbol{\alpha}(s)\boldsymbol{c}(s)$$

$$\mathbf{V}_v(s) = var(\mathbf{v}_{s-1,t}) = \Gamma(0) - \boldsymbol{\beta}(s)\mathbf{b}(s)$$

$$\mathbf{V}_{vu}(s) = cov(\mathbf{v}_{s-1,t}, \mathbf{u}_{s-1,t+s})$$

$$= \Gamma(s) - \mathbf{b}'(s)\boldsymbol{\alpha}'(s)$$

$$\mathbf{u}_{s-1,t} = \mathbf{Z}_{t+s} - \boldsymbol{\alpha}_{s-1,1}\mathbf{Z}_{t+s-1} - \ldots - \boldsymbol{\alpha}_{s-1,s-1}\mathbf{Z}_{t+1}$$

$$\mathbf{v}_{s-1,t} = \mathbf{Z}_t - \boldsymbol{\beta}_{s-1,1}\mathbf{Z}_{t+1} - \ldots - \boldsymbol{\beta}_{s-1,s-1}\mathbf{Z}_{t+s-1}$$

$\mathbf{D}_{\mathrm{v}}(s)$ It is a diagonal matrix with the $i-th$ element being the square root of the $i-th$ diagonal part of $\mathbf{V}_{\mathrm{v}}(s)$ and $\mathbf{D}_u(s)$ is a diagonal matrix with the $i-th$ element being the square root of the $i-th$ diagonal part $\mathbf{V}_u(s)$.

$$\mathbf{b}(s) = \begin{pmatrix} \Gamma'(s-1) \\ \Gamma'(s-2) \\ \vdots \\ \Gamma'(1) \end{pmatrix}, \mathbf{c}(s) = \begin{pmatrix} \Gamma(1) \\ \Gamma(2) \\ \vdots \\ \Gamma'(s-1) \end{pmatrix}$$

$$\boldsymbol{\alpha}'(s) = \begin{pmatrix} \boldsymbol{\alpha}'_{s-1,1} \\ \boldsymbol{\alpha}'_{s-1,2} \\ \vdots \\ \boldsymbol{\alpha}'_{s-1,s-1} \end{pmatrix}, \boldsymbol{\beta}' = \begin{pmatrix} \boldsymbol{\beta}'_{s-1,s-1} \\ \boldsymbol{\beta}'_{s-1,s-2} \\ \vdots \\ \boldsymbol{\beta}'_{s-1,1} \end{pmatrix}$$

Order identification is made by looking at the (+) and (-) signs on the partial correlation value. The symbol (+) is given for the $P_{ij}$ a value which is greater than $2/\sqrt{n}$, the sign (-) for the value smaller than $-2/\sqrt{n}$ and the sign (.) for $P_{ij}$ whose value is $-2/\sqrt{n}$ until $2/\sqrt{n}$ [22].

### 2.2.3 Parameter Estimation

One way to estimate the parameter $\mathbf{\Phi}$ of the VAR model is the method of least squares [22].

$$\mathbf{Z}_t = \mathbf{\Phi}_t + \mathbf{\Phi}_t \mathbf{Z}_{t-1} + \ldots + \mathbf{\Phi}_t \mathbf{Z}_{t-p} + \mathbf{a}_t, t = p+1, \ldots n$$
$$\mathbf{\Phi} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}) \qquad (5)$$

The following are the test statistics used.

$$t_{statistic} = \beta / SE(\beta) \qquad (6)$$

The test statistic is rejected if the value $|t_{statistic}| = t_{\alpha/2,(n-m)}$ is at a significant level of 10%, so it is concluded that the VAR model parameters are significant [25]. The VAR model is appropriate if the residuals meet the white noise assumption and have a multivariate normal distribution.

The white noise test was used to determine that the residuals were independent and the homogeneity of the residuals. White noise test using the Ljung Box test below.

Hypothesis:

$H_0 : \rho_1 = \rho_2 = ... = \rho_m$

$H_1$ : There is at least one $\rho_i$ that is not equal to zero.

Where M is the number of parameters in the model and n is the number of effective observations equal to the number of residuals calculated from the series. Meanwhile, multivariate normal testing was carried out using Henze-Zirkler's test with the following test statistics [26]:

$$HZ = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}D_{ij}} - 2(1+\beta^2)^{-p/2}\frac{1}{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}D_{ij}} + (1+2\beta^2)^{-p/2}$$

$$= \beta = \frac{1}{\sqrt{2}}\left(\frac{n(2p+1)}{4}\right)^{1/p+4} \qquad (7)$$

$$D_{ij} = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T S^{-1}(\mathbf{x}_i - \mathbf{x}_j)$$

$$D_{ij} = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T S^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

The result $p-value$ is greater than the specified significance level, and it can be concluded that the data is normally distributed in multivariate.

## 2.3 $T^2$ Hotelling Chart Based On Residuals

The $T^2$ Hotelling control chart is one of the popular multivariate process control for monitoring the mean vector of the process. the control chart is a development of the univariate control chart $\overline{\text{X}}$ Shewhart. There are two versions of the $T^2$ Hotelling control chart: individual and subgroup. This study used a control chart $T^2$ Hotelling for individuals [27].

Suppose given the number of subgroups $n = 1$, $m$ is the number of observations on each subset, while $p$ is the sum characteristic of the observed control process. If $\mathbf{x_i}$, where $i = 1, 2, ..., m$ is the autocorrelated process data. Then the time series model of quality characteristics is as follows.

$$\hat{Z}_t = \xi + \phi Z_{t-1} + \varepsilon_t \qquad (8)$$

It $\hat{Z}_t$ is the estimated value of $Z_t$, then the residual value of $\hat{e}_t$ can be calculated by the following equation:

$$\hat{e}_t = Z_t - \hat{Z}_t \qquad (9)$$

In this case, the $T^2$ Hotelling control chart for $\hat{e}_t$ can be expressed as

$$T_i^2 = \hat{e}_t' \hat{\Sigma}_{et}^{-1} \hat{e}_t \sim \chi^2 \qquad (10)$$

Where $\hat{\Sigma}_{e_t}$ is the variance of the covariance matrix obtained from:

$$\hat{\Sigma}_{e_t} = \frac{1}{T}\sum_t^T \hat{e}_t \hat{e}_t' \qquad (11)$$

Moreover, the control limit of $T^2$ Hotelling can be obtained below:

$$UCL = \frac{p(m+1)(m-1)}{m^2 - mp} F_{(\alpha,p,m-p)'} \qquad (12)$$

$$LCL = 0$$

When the number of sample m is large, m>100 can use an approximate control limit, either

$$UCL = \chi^2_{\alpha,p} \qquad (13)$$

### 2.3.1 $T^2$ Hotelling Single Bootstrap Chart

Bootstrap control limit on multivariate data was first introduced by Phaladiganon *et al.* [9]. The bootstrap method is more convenient for establishing control limits as it does not contain any modeling process in specifying the parameters. The steps for calculating $T^2$ PB Chart are as follows:

**Step 1**: Calculating $T^2$ statistics from the residual data using equation (11) and obtain:

$$\left\{T^2_{Var1}, T^2_{Var2}, ..., T^2_{Varm}\right\} \qquad (14)$$

**Step 2**: Resampling bootstrap B times. The notation (*) indicates the first resampling result

$$\left\{\begin{bmatrix} T_1^{2*(1)} & T_2^{2*(1)} & \cdots & T_m^{2*(1)} \\ T_1^{2*(2)} & T_1^{2*(2)} & \cdots & T_m^{2*(2)} \\ \vdots & \vdots & \cdots & \vdots \\ T_1^{2*(B)} & T_1^{2*(B)} & \cdots & T_m^{2*(B)} \end{bmatrix}\right\}$$

**Step 3**: Determine the value $T^2$ statistic based on the $100(1-\alpha)$ percentile for each bootstrap sample.

$$\left\{\begin{bmatrix} T^{2*(1)}_{100(1-\alpha)} \\ T^{2*(2)}_{100(1-\alpha)} \\ \vdots \\ T^{2*(B)}_{100(1-\alpha)} \end{bmatrix}\right\}$$

**Step 4**: Calculate the control limit based on the average statistical values $T^2$ $100(1-\alpha)$ percentile.

$$UCL_{PB} = \frac{1}{B}\sum_{i=1}^{B} T^{2*(i)}_{100(1-\alpha)} = \overline{T}^{2*(i)}_{100(1-\alpha)} \pm 3\sigma \quad (15)$$

Then Mostajeran *et al.* introduced $T^2$ NB Chart for calculating the control limit for $T^2$ Hotelling by resampling the original data first, not on the $T^2$ statistics that have been built [19]. The steps on the $T^2$ NB Chart are shown as follows.

**Step 1**: Suppose given $\hat{\boldsymbol{e}} = \left[\hat{e}_1, \hat{e}_2, ..., \hat{e}_m\right]$ from the sample $\mathbf{X}$, generate Bootstrap $\hat{\boldsymbol{e}}$ with replacement as $B_1$ times and obtained $\mathrm{X}^* = \left[\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_B^*\right]$**.**

The notation (*) indicates the first resampling result.

**Step 2**: Calculate the $T^2$ statistic with equation (3) in each bootstrap sample.

$T_1^{2*}, T_2^{2*}, \cdots, T_B^{2*}$.

**Step 3**: Determine $B(1-\alpha)th$ percentile values $T^{2*}$ as the upper control limit.

$$UCL_{NB} = T^{2*}_{[B(1-\alpha]} \pm 3\sigma \qquad (16)$$

### 2.3.2 $T^2$ Hotelling Double Bootstrap Chart

The double bootstrap (DB) procedure is the doctrine of generating new data from the bootstrap data set that has been developed previously. From the first-stage bootstrap data set $B_1$ replicated from the original data set, The bootstrap process is repeated for $B_2$ replications so that the total number of test statistics must be calculated $B_1 + B_1B_2$ [28]. The concept was adapted as one new algorithm proposed to establish sensitive control limits on all shifts and could also be used on all observation measures. Our proposed control chart builds on the control limits introduced by Phaladiganon *et al.* [9] and Mostajeran *et al.* [19]. Such as the theory presented by Efron [11], Bootstrap is generated first on the original data as $B_1$ times. Then resampling was carried out again on the $T^2$ statistics built from the first resampling $B_2$ times. So the control limit calculation process is as much as $B_1 + B_1B_2$ time. The proposed algorithm is summarized as follows:

**Step 1**: Suppose given data $\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\right]$ with autocorrelation. Then calculate the residual $\hat{e}_t$ using equation (9).

**Step 2**: Generate bootstrap sample from residual in step 1 with replacement as $B_1$ times. The bootstrap sample that has been replicated is shown in the following matrix:

$$\hat{e}_i^{*(t)} = \begin{bmatrix} \hat{e}_1^{*(1)} & \hat{e}_1^{*(2)} & \cdots & \hat{e}_1^{*(B_1)} \\ \hat{e}_2^{*(1)} & \hat{e}_2^{*(2)} & \cdots & \hat{e}_2^{*(B_1)} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{e}_m^{*(1)} & \hat{e}_m^{*(2)} & \cdots & \hat{e}_m^{*(B_1)} \end{bmatrix}$$

The notation (*) indicates the single bootstrap replication result.

**Step 3**: Calculate $T^2$ statistics with equation (3) in each bootstrap sample.

$T_1^{2*}, T_2^{2*}, \cdots, T_{B_1}^{2*}$ .

**Step 3:** Generate a second bootstrapping on each $T^{2*}$ as $B_2$ times, and obtain:

$$\left\{ \begin{matrix} T_1^{2**(1)} & T_2^{2**(1)} & \cdots & T_{B_1}^{2**(1)} \\ T_1^{2**(1)} & T_2^{2**(1)} & \cdots & T_{B_1}^{2**(1)} \\ \vdots & \vdots & \cdots & \vdots \\ T_1^{2**(1)} & T_2^{2**(B_2)} & \cdots & T_{B_1}^{2**(B_2)} \end{matrix} \right\}$$

**Step 4**: Determine the value $T^{2**}$ statistic based on the $100(1-\alpha)$ percentile for each double bootstrap sample

$$\left\{ \begin{matrix} T_{100(1-\alpha)}^{2**(1)} \\ T_{100(1-\alpha)}^{2**(1)} \\ \vdots \\ T_{100(1-\alpha)}^{2**(B_2)} \end{matrix} \right\}$$

**Step 5**: Calculate the control limit based on the average of the statistical values $T^{2**}$ 100(1-α) percentile

$$UCL_{DB} = \frac{1}{B_2}\sum_{i=1}^{B_2} T_{100(1-\alpha)}^{2**(i)} = \overline{T}_{100(1-\alpha)}^{2**(i)} \pm 3\sigma \quad (17)$$

## 2.4 Average Run Lenght

After obtaining the control limit in the previous step, the next step calculates $T^2$ statistics from the sample $\hat{e}_t$ using equation (3). Then The value of the $T^2$ statistic is plotted in the $T^2$ control chart. If $T^2{}_i > UCL_{DB}$ The sample is classified as conforming, we move back to point 1. Otherwise, if $T^2{}_i < UCL_{DB}$ The sample is classified as non-conforming

ARL is the average time plotting the points on the control graph before an out-of-control issue is detected [27]. The value of ARL for a given shift of magnitude $\delta$ can be obtained as follows [29]:

$$ARL(\delta) = E\left[ARL_{ConformingRunLenght}\right] * E\left[ConformingRunLenght\right] \quad (18)$$
$$= \frac{1}{1-(1-q)^L} * \frac{1}{q}$$

Where q is the probability of a sample being non-conforming.

## 3. RESULT AND DISCUSSION

The data used in this study is air quality data from one month in the air monitoring station area of Tandes, Surabaya, East Java, Indonesia. The air quality analyzed in this study consisted of three quality characteristics, Particulate Matter 10 (PM 10), Particulate Matter 2.5 (PM 2.5), and Carbon Monoxide (CO). The three characteristics have a strong correlation ( $\rho = 0.78$ ) for the characteristics of PM 2.5 and PM 10, ( $\rho = 0.41$ ) for the characteristics of PM 10 and CO, as well as CO and PM 2.5 correlate ( $\rho = 0.50$ ). In addition, it is multivariate with equation (1) showing means Reject H0 and concluding that the quality characteristics are correlated with each other. Therefore, air quality with PM 10, PM 2.5, and CO characteristics can be done simultaneously using the Multivariate $T^2$ Hotelling control chart.

Furthermore, the test is carried out to determine the presence of autocorrelation between observations for each characteristic. The test is carried out by looking at the ACF plot presented in Figure (1). Figure (1) shows that the air quality characteristics have an autocorrelation value exceeding the significance limit at lag 1 to lag 24 for PM 10. Then lag 1 to 23 for PM 2.5 and lag 1 to 29 for CO. Therefore, using a time series control chart is the right choice for air quality control in Surabaya.

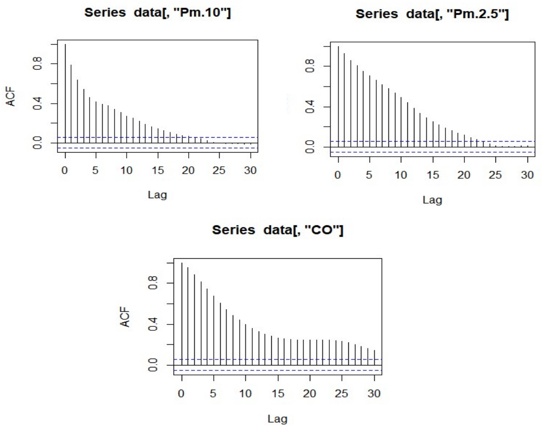

*Figure 1: Plot ACF*

### 3.1 Vector Autoregressive (VAR) Modelling

Before doing time series modeling using VAR, stationarity check for variance and the mean are done first. Suppose the quality characteristics have been stationary in the mean and variance. The next step is identifying the optimum VAR orde by calculating the partial cross-correlation value using equation 15. It is shown that the partial cross-correlation value is significant at lags 1, 2, 3, 4, 5, 6, 7, 11, and 12. To clarify, determining the optimum order VAR is done by looking at Akaike's Information Criterion (AIC) value. The model is selected based on the smallest AIC value, as shown in Table 1.

*Table 1: Information criterion models*

| Lag | AIC value |
|---|---|
| 1 | -27.38572 |
| 2 | -27.52133 |
| 3 | -27.54024 |
| 4 | -27.53637 |
| 5 | -27.53664 |
| 6 | -27.53688 |
| 7 | -27.53750 |

The smallest AIC value based on the table is in lag 3, so the most optimum estimation model is a VAR with ordo 3. Furthermore, parameter estimation can be carried out.

After restricting to eliminate insignificant parameters, we get a model with the parameters in Table 4 below.

*Table 4: Parameter Estimation*

| Output Variable | Parameter | Estimated Value | *p-value* |
|---|---|---|---|
| PM 10 | $\phi_{111}$ | 0.8587042 | $< 0.001$ |
| | $\phi_{313}$ | -0.2819557 | $< 0.001$ |
| | $\phi_{314}$ | 3.084 | $< 0.05$ |
| | $\phi_{211}$ | 0.19710 | $< 0.01$ |
| PM 2.5 | $\phi_{221}$ | 0.89727 | $< 0.001$ |
| | $\phi_{122}$ | 0.0677584 | $< 0.05$ |
| | $\phi_{322}$ | -0.14073 | $< 0.01$ |
| | $\phi_{323}$ | 0.0130546 | $< 0.01$ |
| | $\phi_{233}$ | -0.0177211 | $< 0.05$ |
| CO | $\phi_{131}$ | 1.2170102 | $< 0.001$ |
| | $\phi_{232}$ | 0.55930 | $< 0.01$ |
| | $\phi_{333}$ | -0.2989686 | $< 0.001$ |

After all significant parameters were obtained, diagnostic checking was carried out on the residual model. The residuals of the VAR model must meet the white noise and multivariate normal distribution assumptions. From the results of the white noise test using the Ljung Box test, it was found that the *p-value* of the test was 0.9665 for the PM 10 characteristic, 0.9708 for the PM 2.5, and 0.8733 for the CO characteristic. So that it is possible to be concluded that the *p-values* of the three are greater than $\alpha = 0.05$, which means reject $H_0$ , and it can be concluded that the data values are independent.

Meanwhile, for testing the normality of the residual model with equation (7), the *p-value* is smaller than 0.05. It can be concluded that the data is non-normally distributed in multivariate [26]. Therefore, the bootstrap method for the $T^2$ Hotelling control chart is recommended to be carried out in the next step.

### 3.2 $T^2_{VAR}$ Chart Based On Residual

After obtaining the residuals from the previously formed VAR model, the next step is to create a control chart with Bootstrap control limits. The performance of the proposed control limits is evaluated, and the ARL values are compared with $T^2_{VAR}$ NB chart and the $T^2_{VAR}$ PB chart at different observations. Daily air quality data indicate the number of small observations, N=49, while the weekly air quality data is N=338, and the number of large observations with monthly data is N=1394. Then the value of ARL is also compared with the given shift ($\delta$) in the average vector with small $(0.001 \le \delta \le 0.009)$, moderate $(0.1 \le \delta \le 0.9)$, and large shifts $(1 \le \delta \le 3)$. A comparison of ARL values using selection shift ($\delta$) for N=49 is shown in Table 5 below.

*Table 5: Comparison of ARL values using selection shift ($\delta$) for N=49*

| N | Comparison of ARL Value | | | |
|---|---|---|---|---|
| | Shift ($\delta$) | $T^2$ Nb Chart | $T^2$ Pb Chart | $T^2$ DB Chart |
| 49 | 0.001 | 90.8685 | 24.2301 | 18.1936 |
| | 0.002 | 90.8411 | 24.2236 | 18.1889 |
| | 0.003 | 90.8137 | 24.2170 | 18.1842 |
| | 0.004 | 90.7863 | 24.2104 | 18.1794 |
| | 0.005 | 90.7589 | 24.2039 | 18.1747 |
| | 0.009 | 90.6496 | 24.1777 | 18.1558 |
| | 0.1 | 88.2388 | 23.6013 | 17.7395 |
| | 0.3 | 83.4092 | 22.4477 | 16.9068 |
| | 0.5 | 79.1335 | 21.4279 | 16.1713 |
| | 0.7 | 75.3219 | 20.5202 | 15.5172 |
| | 0.9 | 71.9029 | 19.7074 | 14.9320 |
| | 1 | 70.3223 | 19.3322 | 14.6620 |
| | 1.3 | 66.0240 | 18.3135 | 13.9297 |
| | 1.5 | 63.4787 | 17.7118 | 13.4976 |
| | 1.7 | 61.1515 | 17.1627 | 13.1038 |
| | 1.9 | 59.0156 | 16.6599 | 12.7435 |
| | 2 | 58.0123 | 16.4240 | 12.5747 |
| | 2.1 | 57.0487 | 16.1978 | 12.4128 |
| | 2.3 | 55.2314 | 15.7719 | 12.1084 |
| | 2.5 | 53.5476 | 15.3782 | 11.8273 |
| | 2.7 | 51.9831 | 15.0133 | 11.5671 |
| | 3 | 49.8340 | 14.5137 | 11.2114 |

Table 5 shows that the ARL values for small, medium and large shifts decreased with increased shifts. It can also be seen that the ARL value for the Proposed $T^2_{VAR}$ DB chart has the smallest ARL value compared to the $T^2_{VAR}$ NB chart and $T^2_{VAR}$ PB chart for all shifts. It shows that the proposed $T^2_{VAR}$ DB chart is faster at detecting out-of-control points. The Proposed $T^2_{VAR}$ DB chart can detect an out-of-control point at a shift of 0.001, with 18 observations. The $T^2_{VAR}$ NB and PB charts require 90 and 24 observations to detect out-of-control points.

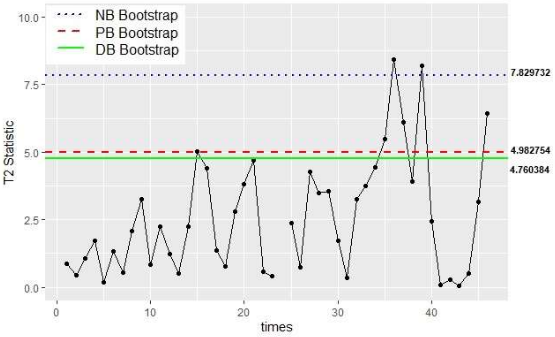

*Figure 2: Control limit established by the $T^2_{VAR}$ NB chart, $T^2_{VAR}$ NB chart, and proposed $T^2_{VAR}$ NB chart with 49 observations.*

The sensitivity of the Proposed $T^2_{VAR}$ DB chart is also shown in Figure 2. With a UCL value of 4.760384, the Proposed $T^2_{VAR}$ DB chart can detect seven out-of-control points, which is more sensitive than the $T^2_{VAR}$ NB chart, which only detects three out-of-control points with a UCL value of 7.829732. Similarly, the $T^2_{VAR}$ PB chart can detect six out-of-control points with a UCL value of 4.962754. Then a comparison of ARL values using selection shift ($\delta$) for N=338 is shown in Table 6 below.

*Table 6: Comparison of ARL estimation using selection shift ( $\delta$ ) for N=338*

| N | Comparison of ARL Value | | | |
|---|---|---|---|---|
| | Shift ( $\delta$ ) | $T^2$ Nb Chart | $T^2$ Pb Chart | $T^2$ DB Chart |
| 338 | 0.001 | 46.2871 | 29.5640 | 17.0856 |
| | 0.002 | 46.2738 | 29.5558 | 17.0812 |
| | 0.003 | 46.2604 | 29.5476 | 17.0768 |
| | 0.004 | 46.2471 | 29.5395 | 17.0724 |
| | 0.005 | 46.2337 | 29.5313 | 17.0680 |
| | 0.009 | 46.1805 | 29.4986 | 17.0504 |
| | 0.1 | 45.0058 | 28.7790 | 16.6632 |
| | 0.3 | 42.6533 | 27.3385 | 15.8889 |
| | 0.5 | 40.5717 | 26.0646 | 15.2051 |
| | 0.7 | 38.7170 | 24.9303 | 14.5970 |
| | 0.9 | 37.0543 | 23.9141 | 14.0532 |
| | 1 | 36.2859 | 23.4448 | 13.8023 |
| | 1.3 | 34.1976 | 22.1702 | 13.1220 |
| | 1.5 | 32.9620 | 21.4167 | 12.7208 |
| | 1.7 | 31.8331 | 20.7289 | 12.3553 |
| | 1.9 | 30.7977 | 20.0986 | 12.0209 |
| | 2 | 30.3115 | 19.8029 | 11.8643 |
| | 2.1 | 29.8448 | 19.5191 | 11.7141 |
| | 2.3 | 28.9652 | 18.9846 | 11.4317 |
| | 2.5 | 28.1507 | 18.4902 | 11.1711 |
| | 2.7 | 27.3946 | 18.0317 | 10.9299 |
| | 3 | 26.3571 | 17.4034 | 10.6004 |

Meanwhile, for the moderate number of observations, N = 338, the Proposed $T^2_{VAR}$ DB chart also has good sensitivity to all shifts. It can be shown by Table 6 that the ARL values for all shifts for the Proposed $T^2_{VAR}$ DB chart are smaller than the $T^2_{VAR}$ NB chart and $T^2_{VAR}$ PB chart.

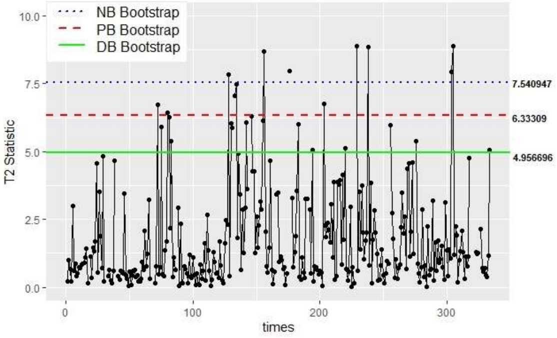

*Figure 3: Control limit established by the $T^2_{VAR}$ NB chart, $T^2_{VAR}$ NB chart, and proposed $T^2_{VAR}$ NB chart with 338 observations.*

Similar to the number of small observations, the Proposed $T^2_{VAR}$ DB chart can also detect more out-of-control points than the two control charts shown in figure 3. Out-of-control points detected by the Proposed $T^2_{VAR}$ DB chart are 51 with a UCL value of 4.805834. It is more than the $T^2_{VAR}$ NB chart, which can detect out-of-control as 36 points, and the $T^2_{VAR}$ DB can detect out-of-control as 31 points. Then a comparison of ARL values using selection shift ( $\delta$ ) for N=1393 is shown in Table 7 below.

*Table 7: Comparison of ARL estimation using selection shift ( $\delta$ ) for N=1393*

| N | Comparison of ARL Value | | | |
|---|---|---|---|---|
| | Shift ( $\delta$ ) | $T^2$ Nb Chart | $T^2$ Pb Chart | $T^2$ DB Chart |
| 1393 | 0.001 | 76.8705 | 76.8705 | 34.6889 |
| | 0.002 | 76.8475 | 76.8475 | 34.6791 |
| | 0.003 | 76.8246 | 76.8246 | 34.6694 |
| | 0.004 | 76.8016 | 76.8016 | 34.6596 |
| | 0.005 | 76.7787 | 76.7787 | 34.6499 |
| | 0.009 | 76.6870 | 76.6870 | 34.6109 |
| | 0.1 | 74.6667 | 74.6667 | 33.7527 |
| | 0.3 | 70.6195 | 70.6195 | 32.0344 |
| | 0.5 | 67.0369 | 67.0369 | 30.5145 |
| | 0.7 | 63.8434 | 63.8434 | 29.1608 |
| | 0.9 | 60.9791 | 60.9791 | 27.9477 |
| | 1 | 59.6551 | 59.6551 | 27.3872 |
| | 1.3 | 56.0548 | 56.0548 | 25.8648 |
| | 1.5 | 53.9232 | 53.9232 | 24.9645 |
| | 1.7 | 51.9745 | 51.9745 | 24.1423 |
| | 1.9 | 50.1862 | 50.1862 | 23.3886 |
| | 2 | 49.3463 | 49.3463 | 23.0349 |
| | 2.1 | 48.5396 | 48.5396 | 22.6954 |
| | 2.3 | 47.0185 | 47.0185 | 22.0558 |
| | 2.5 | 45.6092 | 45.6092 | 21.4639 |
| | 2.7 | 44.3001 | 44.3001 | 20.9148 |
| | 3 | 42.5021 | 42.5021 | 20.1618 |

Similar to the previous analysis, the Proposed $T^2_{VAR}$ DB chart on a large number of observations also shows better sensitivity than the $T^2_{VAR}$ NB and $T^2_{VAR}$ PB charts. In table 7, it is shown that the Proposed $T^2_{VAR}$ DB chart has excellent and stable values on small to large shifts.

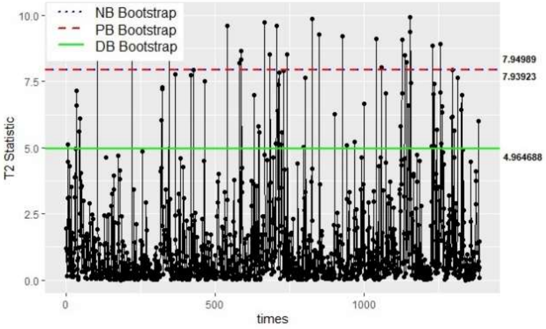

*Figure 4: Control limit established by the $T^2_{VAR}$ NB chart, $T^2_{VAR}$ NB chart, and proposed $T^2_{VAR}$ NB chart with 1394 observations.*

Figure 4 shows that for a large number of observations, i.e., N is 1393, the Proposed $T^2_{VAR}$ DB chart can detect 146 out-of-control points with a UCL of 4.964688. while the out-of-control points that the $T^2_{VAR}$ NB chart can detect are fewer, that is, 96 points. The $T^2_{VAR}$ PB chart detects an out-of-control point which is not much different from the $T^2_{VAR}$ NB chart, which is 97 points with a UCL value of 7.93932.

## 4. CONCLUSIONS

In the multivariate process using $T^2$ Hotelling, autocorrelation must have violated the assumptions of the control chart. Autoregressive vectors estimate and monitor residual VAR as an independent multivariate serial series. Choosing the correct control limit can make the process control more accurate. This study proposed finding a sensitive control chart to monitor shifts in data with autocorrelation using $T^2$ Hotelling of the VAR residual model and Double Bootstrap to build a control limit ( $T^2_{VAR}$ DB Chart).

The sensitivity of the control chart to process shifts can be shown by the number of out-of-control points detected. This study found that the proposed chart detects more out-of-control points than the PB and NB charts. The sensitivity was tested on several different observations. Evaluation of ARL value with a shift in the chart also results that the proposed $T^2_{VAR}$ DB chart can detect out-of-control faster than $T^2_{VAR}$ PB chart dan $T^2_{VAR}$ NB Chart. Without a shift, the proposed $T^2_{VAR}$ DB chart requires 22 observations, $T^2_{VAR}$ PB chart 85 observations, and $T^2_{VAR}$ NB chart requires 46 observations. The ARL value indicates the amount of data evaluated before the out-of-control point was detected [27]. Faster detection of out-of-control points can minimize false alarms in out-of-control detection.

Thus, by applying air quality control data in Surabaya as an individual observation. We conclude that the scheme proposed for monitoring shifts in the process average with the residual VAR approach and the Double Bootstrap method can improve the $T^2$ Hotelling control chart to monitor multivariate processes with autocorrelation. After obtaining the out-of-control point, researchers can analyze the assignable cause of the uncontrolled air quality so that a solution is obtained from the uncontrolled air quality in Surabaya. Further research can develop the use of other estimators in measuring. Parameter estimation model VAR using other parameters estimator model and combined with the determination of control limits of the Autoregressive Model that have been proposed can be developed in future research.

## ACKNOWLEDGEMENT

We thank to the editor and referees for their valuable comments, which significantly improved the quality of the articles. This work is supported by Hibah Penelitian Tesis Magister Direktorat Pendidikan Tinggi (DIKTI) - Republik Indonesia.

## REFERENCES:

[1] Air Quality Life Index, “Air Quality Life Index,” *aqli.epic.uchicago.edu*, 2019. https://aqli.epic.uchicago.edu/country-spotlight/indonesia/ (accessed Jan. 10, 2020).

[2] World Health Organization, “Air pollution,” *www.who.int*, 2022. https://www.who.int/health-topics/air-pollution#tab=tab_1 (accessed Jan. 10, 2022).

[3] R. C. Leoni, A. F. B. Costa, and M. A. G. Machado, “The effect of the autocorrelation on the performance of the T2 chart,” *Eur. J. Oper. Res.*, vol. 247, no. 1, 2015, pp. 155–165.

[4] W. Kliengchuay, S. Worakhunpiset, Y. Limpanont, A. C. Meeyai, and K. Tantrakarnapa, “Influence of the meteorological conditions and some pollutants on PM10 concentrations in Lamphun, Thailand,” *J. Environ. Heal. Sci. Eng.*, vol. 19, no. 1, 2021, pp. 237–249.

[5] C. Marchant, V. Leiva, F. J. A. Cysneiros, and J. F. Vivanco, "Diagnostics in multivariate generalized Birnbaum-Saunders regression models," *J. Appl. Stat.*, vol. 43, no. 15, 2016, pp. 2829–2849.

[6] R. J. Tibshirani and B. Efron, "An introduction to the bootstrap," *Monogr. Stat. Appl. Probab.*, vol. 57, 1993, pp. 1–436.

[7] L. C. Alwan and H. V Roberts, "Time-series modeling for statistical process control," *J. Bus. Econ. Stat.*, vol. 6, no. 1, 1988, pp. 87–95.

[8] X. Pan and J. Jarrett, "Using vector autoregressive residuals to monitor multivariate processes in the presence of serial correlation," *Int. J. Prod. Econ.*, vol. 106, no. 1, 2007, pp. 204–216.

[9] P. Phaladiganon, S. B. Kim, V. C. P. Chen, J.-G. Baek, and S.-K. Park, "Bootstrap-based T 2 multivariate control charts," *Commun. Stat. Comput.*, vol. 40, no. 5, 2011, pp. 645–662.

[10] E. Vanhatalo and M. Kulahci, "The effect of autocorrelation on the Hotelling T2 control chart," *Qual. Reliab. Eng. Int.*, vol. 31, no. 8, 2015, pp. 1779–1796.

[11] B. Efron, "Bootstrap methods: another look at the jackknife annals of statistics 7: 1–26," *View Artic. PubMed/NCBI Google Sch.*, vol. 24, 1979.

[12] S. M. Bajgier, "The use of bootstrap to construct limits on control charts," in *Proceedings of the decision Science Institute*, 1992, pp. 1611–1613.

[13] M. D. Nichols and W. J. Padgett, "A bootstrap control chart for weibull percentiles," *Qual. Reliab. Eng. Int.*, vol. 22, no. 2, 2006, pp. 141–151.

[14] Y. L. Lio and C. Park, "A bootstrap control chart for Birnbaum-Saunders percentiles," *Qual. Reliab. Eng. Int.*, vol. 24, no. 5, 2008, pp. 585–600.

[15] P. Taylor, Y. L. Lio, and C. Park, "Journal of Statistical Computation and A bootstrap control chart for inverse Gaussian percentiles," no. June 2013, pp. 37–41.

[16] M. Ahsan, M. Mashuri, and H. Khusna, " Intrusion Detection System Using Bootstrap Resampling Approach of $T^2$ Control Chart Based On Successive Difference Covariance Matrix," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 8, 2018.

[17] A. Mostajeran, N. Iranpanah, and R. Noorossana, "An Explanatory Study on the Non-Parametric Multivariate T2 Control Chart," *J. Mod. Appl. Stat. Methods*, vol. 17, no. 1, 2018, pp. 1–27.

[18] L. de Andrade Mairinque, R. Bruno Dutra Pereira, K. Mota Nascimento, C. Henrique Lauro, and L. Cardoso Brandão, "A bootstrap control chart for the availability index," *Int. J. Adv. Manuf. Technol.*, vol. 120, no. 7, 2022, pp. 5151–5161.

[19] A. Mostajeran, N. Iranpanah, and R. Noorossana, "A New Bootstrap Based Algorithm for Hotelling ' s T2 Multivariate Control Chart," vol. 27, no. 3, 2016, pp. 269–278.

[20] R. Beran, "Prepivoting test statistics: a bootstrap view of asymptotic refinements," *J. Am. Stat. Assoc.*, vol. 83, no. 403, 1988, pp. 687–697.

[21] M. S. Lola, N. H. Zainuddin, M. N. A. Ramlee, and H. Sofyan, "Double bootstrap control chart for monitoring sukuk volatility at Bursa Malaysia," *J. Teknol.*, vol. 79, no. 6, 2017.

[22] W. W. S. Wei, "Time series analysis," in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*, 2006.

[23] J. F. Heyse and W. W. S. Wei, "Modelling the advertising-sales relationship through use of multiple time series techniques," *J. Forecast.*, vol. 4, no. 2, 1985, pp. 165–181.

[24] D. F. Morrison, "Multivariate analysis of variance," *Encycl. Biostat.*, vol. 5, 2005.

[25] R. S. Tsay, *An introduction to analysis of financial data with R*. John Wiley & Sons, 2014.

[26] N. Henze and B. Zirkler, "A class of invariant consistent tests for multivariate normality," *Commun. Stat. Methods*, vol. 19, no. 10, 1990, pp. 3595–3617.

[27] D. C. Montgomery, *Introduction to statistical quality control*. John Wiley & Sons, 2007.

[28] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[29] H. J. Huang and F. L. Chen, "A synthetic control chart for monitoring process dispersion with sample standard deviation," *Comput. Ind. Eng.*, vol. 49, no. 2, 2005, pp. 221–240.