Comparing the Performance of Several Multivariate Control Charts Based on Residual of Multioutput Least Square SVR (MLS-SVR) Model in Monitoring Water Production Process

M Ahsan1* and T R Aulia1
1Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia
*Email: muh.ahsan@its.ac.id

Abstract. Water that is used as the basic human need, requires a processing process to get it. Water quality control in Tirtanadi Water Treatment Plant is still univariate, while theoretically the quality characteristics of water quality are correlated and there is also an autocorrelation due to the continuous process. In this study, quality control is performed on three main variables of water quality characteristics, namely acidity (pH), chlorine residual (ppm), and turbidity (NTU) using several multivariate control charts based on Multioutput Least Square Support Vector Regression (MLS-SVR) residuals. MLS-SVR modelling is used to overcome and get rid of autocorrelation. The input results of the MLS-SVR model are specified from the significant lag of the Partial Autocorrelation Function (PACF), which in this study, is the first lag. The results of the MLS-SVR input model and the optimal combination of hyper-parameters produce residual values that have no autocorrelation anymore. The residuals are used to develop the Hotelling’s $T^2$, Multivariate Exponentially Weighted Moving Average (MEWMA), and Multivariate Cumulative Sum (MCUSUM) control charts. In phase I, we found that the processes are statically controlled. Meanwhile, in phase II, the monitoring results show that there are several out-of-control observations.

Keywords: MLS-SVR; multivariate charts; water quality; autocorrelation

1. Introduction
Clean water that is used as a basic human need requires a treatment process to get it. Acidity (pH), chlorine residual, and turbidity are the three main water quality characteristics that are used to monitor water quality. Currently, water quality control in Tirtanadi Water Treatment Plant is carried out univariately for each quality characteristic. Whereas in theory, the three characteristics of water quality are correlated and there is also an autocorrelation because the water treatment and purification process is carried out continuously. Since the three water quality characteristics are correlated, this study proposes using a multivariate chart.

The control chart is a tool to describe a quality characteristic that has been measured from the sample [1]. The multivariate control chart can be characterized into three types, namely Shewhart [2–7], Multivariate Cumulative Sum (MCUSUM) [8–11], and Multivariate Exponentially Weighted Moving Average (MEWMA) [12–15]. The conventional multivariate control chart assumes that the data are mutually independent[1]. Meanwhile, the water quality data shows a serial dependence or autocorrelation that will cause false alarms if not addressed. Autocorrelation is a condition in which successive observations have a relationship [16]. The existence of autocorrelation between the observations leads to type I error and deluding conclusions about the control state of the process.
Two procedures can be employed to monitor data that has autocorrelation. The first is to monitor the autocorrelation data using a conventional control chart with modified control limits. Another method is to develop a control chart based on the residuals. This approach uses a time-series model to data that has a relationship between successive observations. The residuals of the model are used to monitor the process. The residuals obtained are independent, so it is possible to monitor the process using the residual components. In addition to using the time series method, residuals can also be obtained by using the Multioutput Least Square-Support Vector Regression (MLS-SVR) method [17].

In this study, the quality control of the water production process will be carried out using several types of multivariate control charts, namely Hotelling’s $T^2$, MEWMA, and MCUSUM. The results of the three charts will be compared to see which control chart produces the best results. The remaining parts of this paper are composed as follows: Section 2 presents the MLS-SVR algorithms. Section 3 discusses the methodology of this research. Section 4 presents the results as well as discussions. Finally, Section 5 displays conclusions.

2. Multi-output least square-support vector regression (MLS-SVR)

In this section, the MLS-SVR algorithms are presented. The basic Least Square SVR algorithm only learns the mapping from input to a single output. Let $y = [y_i] \in \mathbb{R}^m$, where $i = 1, 2, \ldots, n$ is the sample size, and $j = 1, 2, \ldots, m$ is the number of output variable. The MLS-SVR algorithm uses a kernel function to perform a nonlinear mapping to a higher dimension, $h$ is a higher dimension called Hilbert Space.

MLS-SVR solves the problem by looking for parameters and $W = (w_1, w_2, \ldots, w_m) \in \mathbb{R}^{m \times n}$ $b = (b_1, b_2, \ldots, b_n)^T \in \mathbb{R}^n$ which can minimize the objective function with the following constraint:

$$\text{minimize } J(w_0, V, \Xi) = \frac{1}{2} (w_0^T w_0) + \frac{\gamma_n}{2m} \text{trace} (V^t V) + \frac{\gamma_0}{2} \text{trace} (\Xi^T \Xi),$$

with constrain $Y = Z W + \text{repmat}(b^T, n, 1) + \Xi$.

The optimization in equation 1 can be made into a Lagrange function as follows:

$$L(w_0, V, b, \alpha) = J(w_0, V, \Xi) - \text{trace} \left( (A)^T \left( Z W + \text{repmat}(b^T, n, 1) + \Xi - Y \right) \right),$$

where $A = (a_1, a_2, \ldots, a_n)^T \in \mathbb{R}^{m \times n}$ is a matrix containing the Lagrange multiplier.

Furthermore, to simplify the solution above, the following equation is formulated.

$$\begin{bmatrix} G & 0_{m \times n} \\ 0_{n \times m} & M \\ \end{bmatrix} \begin{bmatrix} b \\ N^T M^{-1} y \end{bmatrix} = \begin{bmatrix} \text{trace} \left( A^T \left( Z W + \text{repmat}(b^T, n, 1) + \Xi - Y \right) \right) \\ \end{bmatrix}.$$

Equation 3 can be solved by simplifying two sets of linear equations, each of which has a positive definite $M$ matrix so that the MLS-SVR decision function can be obtained as follows:

$$f(x) = \phi(x)^T W + \hat{b} = \phi(x)^T \text{repmat}(\hat{w}, 1, m) + \phi(x)^T \hat{V} + \hat{b}$$

$$= \phi(x)^T \text{repmat} \left( \sum_{j=1}^{m} \zeta_j A_j, 1, m \right) + \sum_{j=1}^{m} \phi(x)^T \zeta_j A_j + \hat{b}$$

$$= \text{repmat} \left( \sum_{j=1}^{m} \sum_{i=1}^{n} \hat{a}_{ij} K(x, x_i), 1, m \right) + \sum_{j=1}^{m} \sum_{i=1}^{n} \hat{a}_{ij} K(x, x_i) + \hat{b}.$$

In this study, the Radial Basis Function (RBF) kernel function is used which can be written as follows:

$$K(x, x_i) = \exp \left( -\sigma \| x - x_i \|^2 \right),$$

with $\sigma > 0$. The RBF kernel function was chosen because of the ease and efficiency of computation time. In the next step, the optimal hyper-parameters $(\gamma', \gamma''$, $\sigma)$ are selected using the grid search method with Mean Square Error (MSE) criterion as follows:
\[\text{MSE}_j = \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \hat{y}_{ij})^2, \quad j = 1, 2, \ldots, m, \quad (6)\]

3. Methodology

3.1. Dataset

The data analyzed are obtained from water samples processed at the Tirtanadi Water Treatment Plant from August 1, 2020, to January 24, 2021. The data used is daily data on the reservoir water quality and will be split into two phases. Phase I is on August 1 – October 31, 2020, and phase II is on November 1, 2020 – January 24, 2021.

3.2. Data structure and variables

The data structure of this research is presented in Table 1. Meanwhile, the specification from each quality characteristic is tabulated in Table 2.

| Observations | Quality characteristics |
|--------------|-------------------------|
|              | Acidity (Y₁)            |
|              | Chlorine Residual (Y₂)  |
|              | Turbidity (Y₃)          |
| 1            | Y₁,₁                    |
| 2            | Y₁,₂                    |
| ...          | ...                     |
| n            | Yₙ,₁                    |

Tabel 1. Data structure

| Notation | Variable                  | specification |
|----------|---------------------------|---------------|
| Y₁       | Acidity (pH)              | 6.7 – 7.2     |
| Y₂       | Chlorine Residual (ppm)   | 0.2 – 1.0     |
| Y₃       | Turbidity (Ntu)           | 0.2 – 2.0     |

Tabel 2. Quality characteristics and their specification

3.3. Research steps

The steps taken in this research are defined as follows:
1. Formulate the problem and determine the research objectives.
2. Collect and input the water quality data.
3. Find the optimal hyperparameter from the MLS-SVR model in phase I.
4. Calculate the residual of the MLS-SVR model from phase I.
5. Calculate the residual of the MLS-SVR model from phase II using the optimal hyperparameter in phase I.
6. Monitor the residual of water quality data using three control charts, namely Hotelling’s $T^2$, MEWMA, and MCUSUM.
7. Compare the performance of the three charts.
8. Draw conclusions and suggestions.

4. Result and Discussion

4.1. Descriptive statistics

In this subsection, the descriptive statistics from phases I dan II are presented. The descriptive statistics of the quality of production water in Phase I are displayed in Table 3. Meanwhile, Table 4 presents the descriptive statistics for phase II. From Table 3-4, it can be seen that the mean from each variable is meeting the specification. Therefore, by looking at these results we can say that the processes are meeting the requirements. However, we need more justification using a control chart whether the processes are in-control or there are out-of-control observations.
### Tabel 3. Descriptive statistics of the quality of production water in Phase I

| Variables          | Mean   | Var    | Min   | Median | Max   |
|--------------------|--------|--------|-------|--------|-------|
| Acidity            | 6.9239 | 0.0106 | 6.58  | 6.95   | 7.10  |
| Chlorine Residual  | 0.4544 | 0.0017 | 0.39  | 0.45   | 0.54  |
| Turbidity          | 1.4057 | 0.0316 | 0.95  | 1.40   | 1.72  |

### Tabel 4. Descriptive statistics of the quality of production water in Phase II

| Variables          | Mean   | Var    | Min   | Median | Max   |
|--------------------|--------|--------|-------|--------|-------|
| Acidity            | 6.8196 | 0.0151 | 6.23  | 6.85   | 7.00  |
| Chlorine Residual  | 0.4731 | 0.0025 | 0.33  | 0.46   | 0.65  |
| Turbidity          | 1.6409 | 0.2950 | 0.87  | 1.54   | 4.43  |

4.2. Monitoring results for phase I

In this subsection, the monitoring results for phase I of water quality data are presented. The phase I data was taken from August 1 – October 31, 2020, and treated as the training data to find the optimal hyperparameter. The optimal hyperparameters are \( \gamma' = 2^{15}, \gamma^* = 2^4, \) and \( \sigma = 2^{-11} \) with an MSE of 0.0107. Figure 1 presents the comparison between the original value and prediction in phase I. According to the figure, it can be seen that the prediction values are almost the same as the original.

![Figure 1](image-url)

**Figure 1.** Comparison between the original value and prediction in phase I for: a) Acidity, b) chlorine residual, and c) turbidity
Figure 2. Monitoring results for phase I from: a) Hotelling’s $T^2$, b) MCUSUM, and c) MEWMA.

Figure 2 shows the monitoring results for phase I using three types of control charts. The MCUSUM chart is constructed using $k=0.5$, while the MEWMA chart is constructed using $\lambda=0.25$. From the results, it can be seen that all of the control charts did not send any signal out-of-control. This is indicated that phase I of the monitoring process is statistically controlled.

4.3. Monitoring results for phase II

The monitoring results of the phase II control chart are displayed in this subsection. The phase II data was taken from November 1, 2020 – January 24, 2021, and treated as the testing data. From Figure 3 it can be seen that prediction values of some quality characteristics have a difficulty to follow the original data. This can be seen in acidity and turbidity quality characteristics.

Figure 3. Comparison between the original value and prediction in phase II for: a) acidity, b) chlorine residual, and c) turbidity.

Figure 4 shows the monitoring results for phase II using three types of control charts. Similar to phase I, the MCUSUM chart is constructed using $k=0.5$, and the MEWMA chart is constructed using $\lambda=0.25$. From the monitoring results, it can be seen that all of the control charts send out-of-control signals. The summary of the number of out-of-control signals is displayed in Table 5. According to the summary, for phase I, all of the control charts did not find the out-of-control observation. Meanwhile,
for phase II, the MCUSUM chart finds the higher number of out-of-control observations by detecting 12 out-of-control observations.

![Figure 4](image1.png)

**Figure 4.** Monitoring results for phase II from: a) hotelling’s $T^2$, b) MCUSUM, and c) MEWMA

| Phase | Control Charts | Number of out-of-control signals |
|-------|----------------|----------------------------------|
| 1     | Hotelling’s $T^2$ | 0                               |
|       | MCUSUM          | 0                               |
|       | MEWMA           | 0                               |
| 2     | Hotelling’s $T^2$ | 2                               |
|       | MCUSUM          | 12                              |
|       | MEWMA           | 5                               |

### 5. Conclusions
In this research, the water quality data is analyzed using multivariate control charts. These charts are Hotelling’s $T^2$, MCUSUM, and MEWMA. However, due to the autocorrelation on the water quality data, the conventional multivariate charts are not appropriate to use. To handle this issue, the MLS-SVR method is employed to calculate the residual of the water quality data. The residuals are used to construct the Hotelling’s $T^2$, MEWMA, and MCUSUM control charts. The monitoring results show that all of the control charts used did not find any out-of-control signal in phase I. Furthermore, the MCUSUM chart finds more out-of-control observations compared to the other charts in phase II.

### 6. References
[1] Montgomery D 2009 *Introduction to statistical quality control* (New York)
[2] Ahsan M, Mashuri M, Kuswanto H and Prastyo D D 2018 Intrusion Detection System using Multivariate Control Chart Hotelling’s T2 based on PCA *Int. J. Adv. Sci. Eng. Inf. Technol.* 8 1905–11
[3] Mashuri M, Haryono H, Aksioma D F, Wibawati W, Ahsan M and Khusna H 2019 Tr (R2) control charts based on kernel density estimation for monitoring multivariate variability process *Cogent Eng.* 6 1665949
[4] Ahsan M, Mashuri M, Kuswanto H, Prastyo D D and Khusna H 2018 T2 Control Chart based on Successive Difference Covariance Matrix for Intrusion Detection System *Journal of Physics: Conference Series* vol 1028 (IOP Publishing) p 12220
[5] Ahsan M, Mashuri M and Khusna H 2018 Intrusion Detection System Using Bootstrap Resampling Approach Of T2 Control Chart Based On Successive Difference Covariance Matrix *J. Theor. Appl. Inf. Technol.* 96 2128–38
[6] Ahsan M, Mashuri M, Lee M H, Kuswanto H and Prastyo D D 2020 Robust adaptive multivariate Hotelling’s T2 control chart based on kernel density estimation for intrusion detection system
[7] Mashuri M, Ahsan M, Lee M H and Prastyo D D 2021 PCA-based Hotelling’s T2 chart with Fast Minimum Covariance Determinant (FMCD) Estimator and Kernel Density Estimation (KDE) for Network Intrusion Detection Comput. Ind. Eng. 107447
[8] Woodall W and Ncube M 1985 Multivariate CUSUM Quality-Control Procedures Technometrics 27 285–92
[9] Khusna H, Mashuri M, Ahsan M, Suhartono S and Prastyo D D 2018 Bootstrap Based Maximum Multivariate CUSUM Control Chart Qual. Technol. Quant. Manag.
[10] Khusna H, Mashuri M, Suhartono, Prastyo D D, Lee M H and Ahsan M 2019 Residual-based maximum MCUSUM control chart for joint monitoring the mean and variability of multivariate autocorrelated processes Prod. Manuf. Res. 7 364–94
[11] Arkat J, Niaki S T A and Abbasi B 2007 Artificial neural networks in applying MCUSUM residuals charts for AR(1) processes Appl. Math. Comput. 189 1889–901
[12] Haq A and Khoo M B C 2019 An adaptive multivariate EWMA chart Comput. Ind. Eng. 127 549–57
[13] Khusna H, Mashuri M, Suhartono S, Prastyo D D and Ahsan M 2018 Multioutput least square SVR based multivariate EWMA control chart: The performance evaluation and application Cogent Eng.
[14] Mahmoud M A and Zahran A R 2010 A multivariate adaptive exponentially weighted moving average control chart Commun. Stat. - Theory Methods 39 606–25
[15] Zaidi F S, Castagliola P, Tran K P and Khoo M B C 2020 Performance of the MEWMA-CoDa control chart in the presence of measurement errors Qual. Reliab. Eng. Int.
[16] Karaoğlan A D 2010 Regression control chart for autocorrelated data
[17] Xu S, An X, Qiao X, Zhu L and Li L 2013 Multi-output least-squares support vector regression machines Pattern Recognit. Lett. 34 1078–84