Quality Control of Water Production Process Using Multivariate Control Charts

Y Trimardiani¹, Wibawati¹*, M S Akbar¹, Suhartono¹ and D D Prastyo¹

¹Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*Corresponding author: wibawati@statistika.its.ac.id

Abstract. Water is very important for life, no life in the world can exist without the availability of water. Turbidity, organic substance, and chlorine residual are the main characteristics of water quality. To monitor the quality of a water production process, Multivariate Exponentially Weighted Moving Covariance Matrix (MEWMC) control chart is used for monitoring the stability of the covariance matrix of a process, and Multivariate Exponentially Weighted Moving Average (MEWMA) control chart is used for monitoring the stability of the mean vector of a process. In this study, the optimum λ for the MEWMC and MEWMA control chart for the water production process during the dry season and the rainy season is λ = 0.1. The variability of water production process during both dry and rainy seasons is in control, while the mean of water production process during both dry and rainy seasons is not in control. Based on the analysis of process capability, it is known that the water production process during the dry season and rainy season is capable of producing water with turbidity level that meets the specification limit, but is not capable to produce water with organic substance and chlorine residual that matches the specification limit.

1. Introduction

Water is very important for life because no life in the world can exist without the availability of water. The water production process in a water treatment plant consists of physical, chemical, and biological treatment. Physical treatment is done by sedimentation and filtration, chemical treatment is done by aeration, coagulation, and flocculation, and biological treatment is done by disinfection. Besides, to maintain the quality of clean water that will be distributed to customers, the water treatment plant conducts water quality testing regularly in the laboratory.

Statistical quality control is the use of statistical methods in monitoring and maintaining the quality of products and services, while Statistical Process Control (SPC) is defined as the use of statistical techniques to control a process or production method. One tool that can be used to control the quality of the process is the control chart [1]. Water quality control is carried out by looking at several quality characteristics, including turbidity, organic substance, and chlorine residual. Because these three quality characteristics are correlated, quality control is carried out with a multivariate control chart.

The multivariate control chart that is sensitive to small shifts in mean processes is the Multivariate Exponentially Weighted Moving Average (MEWMA) control chart proposed by Lowry et al in 1992. The MEWMA control chart is one of the best control charts for detecting shifting average vectors. The shift can also occur for the variability of correlated multivariate characteristics. Hawkins and Maboudou-Tchao in 2008 examined the control chart used for monitoring the shift in variability using the Multivariate Exponentially Weighted Moving Covariance Matrix (MEWMC) control chart.
Water quality control for the dry season and the rainy season is different because there are different treatments during the dry season and rainy season. The turbidity level of raw water during the dry season is around 20-50 NTU, while the turbidity of raw water in the rainy season can reach 500-1000 NTU, so there is a difference in the water production process, which is the addition of alum at the beginning of the process. In this research, the MEWMC control chart is employed to monitor the variability of the water production process while the MEWMA control chart is employed to monitor the average water production process.

2. Literature
This section explains the literature related to the method used for monitoring the water production process, which is as follows.

2.1. Shapiro-Wilk’s Test for Multivariate Normality
The multivariate normal density is a generalization of the univariate normal density to $p \geq 2$ dimensions [2]. The probability density function of the multivariate normal distribution is according to equation (1).

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$ (1)

The $p \times 1$ vector $\mu$ represents the expected value of the random vector $x$, and the $p \times p$ matrix is the variance-covariance matrix of $x$. Multivariate normal distribution testing can be done with Shapiro-Wilk’s test as follows.

Hypothesis:
$H_0$: The data is a sample from the multivariate normal distribution
$H_1$: The data is not a sample from the multivariate normal distribution

Test statistic:
$$W^* = \frac{1}{p} \sum_{j=1}^{p} W_{z_j},$$ (2)

where $W_{z_j}$ is Shapiro-Wilk’s statistic evaluated on the $i$-th coordinate of the transformed observations $Z_{j1}, \ldots, Z_{jp}$, $j = 1, \ldots, p$ [3]. The test based on $W^*$ rejects $H_0$ at a test size $\alpha$ if $W^* < c_{\alpha,n,p}$ or $p$-value < $\alpha$.

2.2. Bartlett Test
The Bartlett sphericity test aims to determine whether the variance is homogeneous between variables in multivariate cases. If the variables $X_1, X_2, \ldots, X_p$ are independent of each other, then the correlation matrix between variables is the same as the identity matrix [4]. The Bartlett sphericity test can be stated in the following hypothesis.

H0: $\mathbf{R} = \mathbf{I}$ (there is no correlation between variables)
H1: $\mathbf{R} \neq \mathbf{I}$ (there is a correlation between variables)

Test Statistics:
$$\chi^2 = -\left[n - 1 - \frac{2p + 5}{6}\right] \ln |\mathbf{R}|,$$ (3)

where $n$ is the number of observations, while $p$ is the number of variables, and $\mathbf{R}$ is the correlation matrix of each variable and $\chi^2\left(\frac{1}{2} p (p-1)\right)$ is the value of the chi-square distribution. $H_0$ is rejected if $\chi^2 > \chi^2\left(\frac{1}{2} p (p-1)\right)$ or if $p$-value < $\alpha$. It can be concluded that the correlation matrix is not the same as the identity matrix or there is a correlation between variables.

2.3. Multivariate Exponentially Weighted Moving Covariance Matrix (MEWMC)
Multivariate Exponentially Weighted Moving Covariance Matrix (MEWMC) control chart is used to monitor the stability of the covariance matrix of a process. It is convenient to work with multistandardized data vectors rather than the “raw” process readings [5]. For this, we find a matrix $A$ with the property $A \Sigma A^T = I_p$ and transform to $u_i = A(x_i - \mu)$. Any matrix $A$ with the needed property will work, but we favor the lower triangular (inverse-Cholesky root) matrix. The chart statistic used in the MEWMC control chart is as follows.

$$c_i = tr(S_i) - \log(S_i) - p,$$

where $p$ is the dimension of the data and $S_i$ is a matrix defined in equation (5).

$$S_0 = I_p,$$

$$S_i = (1 - \lambda)S_{i-1} + \lambda u_i u_i^T,$$

untuk $i=1,2,...,n$.

The chart involves two constants, $\lambda$ and $h$. The smoothing constant $\lambda$ is used to tune the chart to different sizes of change. A small value of $\lambda$ is used if we wish to detect a small shift, whereas a large value of $\lambda$ is used if we are interested to detect large shifts. The control limit $h$ is obtained from the simulation result based on different combinations of $\lambda$, $p$. In-Control Average Run length (IC ARL), and the number of simulations. The MEWMC control chart is formed by plotting $c_i$ against $i$, and signaling a loss of control if $c_i > h$.

2.4. Multivariate Exponentially Weighted Moving Average (MEWMA)

Exponentially Weighted Moving Average (EWMA) control chart is used to detect small shifts in the process mean. In this research, the EWMA control chart is used to detect which variable caused shifts in the MEWMA control chart. EWMA is defined in equation (6).

$$z_i = \lambda x_i + (1 - \lambda)z_{i-1},$$

where $0 < \lambda \leq 1$ and $z_0 = \mu_0$. EWMA control chart is formed by plotting $z_i$ against the sample number $i$. The centerline and control limits for the EWMA control chart are defined in equation (7) [1].

$$UCL = \mu_0 + L\sigma \sqrt{\frac{\lambda}{(2 - \lambda)(1 - (1 - \lambda)^2i)},}$$

$$CL = \mu_0,$$

$$LCL = \mu_0 - L\sigma \sqrt{\frac{\lambda}{(2 - \lambda)(1 - (1 - \lambda)^2i)},}$$

where $UCL$ is the Upper Control Limit, $CL$ is the Center Line, and $LCL$ is the lower control limit. $L$ is the width of the control limits, $\mu_0$ is the process mean, $\sigma$ is the standard deviation, and $\lambda$ is the smoothing constant.

The Multivariate Exponentially Weighted Moving Average (MEWMA) control chart is a control chart that is used to detect small shifts in the process mean. The MEWMA control chart is a generalization of the process for the univariate EWMA defined in the equation (8).

$$z_i = \lambda x_i + (1 - \lambda)z_{i-1},$$

$i = 1,2,...,n$, $x_i$ is the vector of the data for the $i$-th observation, $\lambda$ is the smoothing constant matrix with $\lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ and $z_0 = \mu_0$. The smoothing value used for each variable can be the same or different. If there is no reason for choosing a different smoothing constant for each quality characteristic, then the smoothing constant is $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$ with $0 < \lambda \leq 1$, so the MEWMA vector can be defined in equation (9).

$$z_i = \lambda x_i + (1 - \lambda)z_{i-1},$$

The MEWMA control chart test statistics can be determined using equation (10) [6].

$$T_i^2 = (z_i - \mu_0)^T \Sigma^{-1}_G (z_i - \mu_0),$$
with $\Sigma_x$, according to equation (11).

$$\Sigma_x = \frac{\lambda}{2-\lambda} \left[ 1-(1-\lambda)^2 \right] \Sigma$$  \hspace{1cm} (11)

$\Sigma$ is the covariance matrix of the data. The Upper Control Limit (UCL) value is obtained from the simulation results adjusted for the ARLo value until a convergent UCL value is obtained. MEWMA control charts, like their univariate counterparts, are robust to the assumption of normality [1].

2.5. Process Capability Analysis

Process capability analysis is used to estimate process capability. There is a simple quantitative way of expressing process capabilities. One of the indexes used to measure process capability is $C_p$. The $C_p$ capability index does not consider the mean process against its specification limits. The process capability index which also considers the mean process is $C_{pk}$. There are times when a process only has a one-sided specification limit, either LSL or USL, so it is impossible to calculate $C_p$. However, $C_{pk}$ can still be calculated according to the one-sided specification limit formula, either $C_{pu}$ or $C_{pl}$ [7].

Based on the $3\sigma$ standard, a process is said to be capable if it has a capability index of more than 1 [8]. The capability index suggested by the Automotive Industry Action Group (AIAG) for in-control processes is $C_p$ and $C_{pk}$, while for processes that are not in control, the $P_p$ and $P_{pk}$ process performance indexes are used [1]. The $P_p$ and $P_{pk}$ indices can be calculated according to equation (12).

$$P_p = \frac{USL - LSL}{6s},$$

$$P_{pk} = \min \left( \frac{USL - \mu}{3s}, \frac{\mu - LSL}{3s} \right).$$  \hspace{1cm} (12)

The calculation of the $P_p$ and $P_{pk}$ indices in equation (12) is a calculation for data that meets the normal distribution assumptions. Therefore, in the case of data that is not normally distributed, the calculation of the $P_p$ and $P_{pk}$ indices uses equation (13) [9].

$$P_p = \frac{USL - LSL}{X_{0.99865} - X_{0.00135}},$$

$$P_{pk} = \min \left( \frac{USL - X_{0.5}}{X_{0.99865} - X_{0.5}}, \frac{X_{0.5} - LSL}{X_{0.00135} - X_{0.5}} \right).$$  \hspace{1cm} (13)

In equation (13), $X_{0.99865}$ is the 99.865th percentile of the data, $X_{0.5}$ is the median of the data, and is the $X_{0.00135}$ 0.135th percentile of the data. In the case of multivariate and out of control process, the calculation of process capability is defined in equation (14).

$$MP_p = \sum_{j=1}^{n} W_j P_p (X_j),$$

$$MP_{pk} = \sum_{j=1}^{n} W_j P_{pk} (X_j).$$  \hspace{1cm} (14)

$W_j$ is the weight of the $j$-th variable, provided $\sum_{j=1}^{n} W_j = 1$ [10]. The weight of each variable is determined based on the importance of the quality characteristic variable.

3. Result
This section presented the result of the monitoring water production process.

3.1. Multivariate Normal Test
A multivariate normal test is used to determine whether the data is following multivariate normal distribution or not. The p-value for Shapiro-Wilk’s multivariate normality test for the water production
process during the dry season is $3.806 \times 10^{-14}$ and the water production process during the rainy season is 0.1144. Using $\alpha = 0.05$, the decision taken is to reject $H_0$ for the water production process during the dry season and accept $H_0$ for the production process during the rainy season. It can be concluded that the water production process data during the dry season is not following a multivariate normal distribution, but the water production process data during the rainy season is following a multivariate normal distribution. In this research, we use control charts that are robust to the assumption of multivariate normality. Even if the dry season data is not following a multivariate normal distribution, the analysis is continued.

3.2. Bartlett Test
Bartlett test is used to determine whether there is a correlation between variables used in multivariate cases. In this test, the value $\chi^2$ is 38.447 for water production process data during the dry season and 26.597 for water production process data during the rainy season. The p-value for both tests is 0.000. Using alpha ($\alpha$) = 0.05 and degrees of freedom (df) = 3, the decision is taken is to reject $H_0$. It can be concluded that there is a correlation between the variables used for both the water production process during the dry and rainy season.

3.3. MEWMC Control Charts
The MEWMC control chart is used to monitor process variability. Water quality control for the dry season and the rainy season is different because there are different treatments during the dry season and rainy season.

![Figure 1. MEWMC Control Chart for Dry Season Data using (a) $\lambda = 0.1$ (b) $\lambda = 0.2$, and (c) $\lambda = 0.3$](image)

The MEWMC control chart to monitor the variability of the water production process during the dry season with $\lambda = 0.1, \lambda = 0.2$, and $\lambda = 0.3$ shows that the variability of the water production process during the dry season is not in control, because some observations are out of control limit. For $\lambda = 0.1$, 31 observations are out of control (OOC). For $\lambda = 0.2$, there are 20 OOC observations. For $\lambda = 0.3$, there are 11 OOC observations. After forming the MEWMC control chart, the optimal $\lambda$ is determined. The small value of $\lambda$ is used to see the small process shift, while the large value of $\lambda$ is used to see the large process shift [5]. Determining the optimal $\lambda$ can be done by looking at the number of observations that are OOC. The MEWMC control chart with the highest number of OOC observations shows that the
control chart is more sensitive in detecting shifts in process variability when compared to the MEWMC control chart with a fewer number of OOC observations.

The MEWMC control chart with $\lambda = 0.1$ has the highest number of OOC observations, so $\lambda = 0.1$ was chosen as the optimal smoothing constant because it is more sensitive in detecting small shifts in process variability. Based on the results of discussions with experts, it was found that observations that were out of control were caused by unideal filter conditions, the presence of pollutants, and the unideal dose of disinfectant application. After knowing the cause of the OOC observations, the MEWMC control chart for the dry season is improved by removing the OOC observations. This is done to get in-control variability of the water production process.

Figure 2. In-Control MEWMC Control Chart for Dry Season Data using $\lambda = 0.1$

The MEWMC control chart is improved by removing 13 observations that are outside the control limit. The MEWMC control chart for the dry season is already in control as we can see in Figure 2. Because the variability of the water production process is in control, it can be continued to control the process mean of water production with the MEWMA control chart.

Figure 3. MEWMC Control Chart for Rainy Season Data using (a) $\lambda = 0.1$ (b) $\lambda = 0.2$, and (c) $\lambda = 0.3$

The MEWMC control chart for the water production process during the rainy season shows that none of the observations were out of control. This shows that the variability of the water production process during the rainy season is in control. To determine the optimal value of $\lambda$ for the MEWMC control chart during the rainy season, we use the smallest difference between the maximum value of $c_i$ and the upper
control limit as the criteria. The optimal value of $\lambda$ is 0.1 because it has the smallest difference between the maximum value of $c_i$ and the upper control limit.

3.4. MEWMA Control Charts

After monitoring the process variability using the MEWMC control chart, the MEWMA control chart is formed to detect shifts in process’ mean. The formation of the MEWMA control chart is also separated between the dry season and rainy season because there are differences in conditions and treatment of raw water between the two seasons. In the rainy season, the water discharge tends to be high and there is a lot of sediment in raw water so that the alum used for the production process during the rainy season is a lot more than the alum used for the water production process during the dry season.

![Figure 4. MEWMA Control Chart for Dry Season Data using (a) $\lambda = 0.1$ (b) $\lambda = 0.2$, and (c) $\lambda = 0.3$](image)

Figure 4 shows that there are still many observations that are out of control so that the process mean water production during the dry season is not in control. There are 145 OOC observations for $\lambda = 0.1$, 91 OOC observations for $\lambda = 0.2$ and 57 OOC observations for $\lambda = 0.3$. The optimal $\lambda$ of the MEWMA control chart for the water production process during the dry season is $\lambda = 0.1$ because it produces the most sensitive control chart, $\lambda = 0.1$ is also effective to detect small shifts in the process mean [11]. This can be seen from a large number of OOC observations on the MEWMA control chart with $\lambda = 0.1$.

The process mean of water production during the rainy season is not statistically controlled, as we can see in Figure 5. This is indicated by the number of observations that are outside the control limits on the MEWMA control chart. There are 89 OOC observations in MEWMA control chart with $\lambda = 0.1$, 63 OOC observations in MEWMA control chart with $\lambda = 0.2$, and 41 OOC observations in MEWMA control chart for $\lambda = 0.3$. Determination of the optimal $\lambda$ on the MEWMA control chart for the rainy season is also using the criteria for the highest number of OOC observations. The MEWMA control chart using smaller $\lambda$ is more sensitive in detecting shifts in the process mean. Therefore, the optimal $\lambda$ used to control the process mean of water production during the rainy season with the MEWMA control chart is $\lambda = 0.1$. 
3. Interpretation for Out of Control Observations

After forming the MEWMA control chart, we can see that there are still many observations that are out of control. To detect which variables caused the observation to be out of control, EWMA control charts can be used to detect the shift in the process mean for each variable [11]. Detection of variables causing out of control in the water production process is done by forming an EWMA control chart for each variable with optimum \( \lambda \) that is used to form the MEWMA control chart. Figure 6 shows the EWMA control chart for dry season observations for each variable.

![EWMA Control Chart for Dry Season Data for Each Variables](image)

**Figure 6.** EWMA Control Chart for Dry Season Data for Each Variables (a) Turbidity (b) Organic Substance, and (c) Chlorine Residual
The EWMA control chart for each variable shows that univariately the mean process of water production based on each variable is out of control. Based on the number of observations that are outside the control limits, organic substance and chlorine residual is the main cause for out-of-control observations in the MEWMA control chart because EWMA control charts for these 2 variables have a large number of out of control observations. There is 1 OOC observation for turbidity, 100 OOC observations for organic substance, and 106 OOC observations for chlorine residual.

![EWMA Control Chart](image)

**Figure 7.** EWMA Control Chart for Rainy Season Data for Each Variable (a) Turbidity (b) Organic Substance, and (c) Chlorine Residual

Based on the three EWMA control charts for monitoring the mean process for water production during the rainy season, we can see that univariately the mean process based on the three water quality characteristics variables is not statistically controlled. The organic substance had the highest number of out of control observations, which are 64 OOC observations. While the total of OOC observations for turbidity is 24 and OOC observations for chlorine residual is only 14. This indicates that the organic substance is the main cause of the OOC observations for the MEWMA control chart during the rainy season.

3.6. **Analysis of Process Capability**

Process capability analysis is used to estimate the ability of the water production process to produce water that meets the specification limits. A production process is said to be capable if all observations are within the specification limits. When the observed data were statistically controlled, the process capability indexes used were $C_p$ and $C_{pk}$. However, when the observed data were not statistically controlled, the indexes used were $P_p$ and $P_{pk}$. The $P_p$ and $C_p$ indices cannot be calculated for this study because water quality characteristics have only one-sided specification limits for turbidity and organic substance.

After monitoring the water quality using the MEWMC and MEWMA control charts, it can be seen that the process mean is not statistically controlled for both dry and rainy seasons. So, $P_{pk}$ index is used to calculate the process capability. For multivariate observational data, the calculation of the process capability index considers the weight of each quality characteristic. Each variable has the same
contribution to the quality of the water production process, so the same weight is used for each variable. Table 1 shows the capability index of the water production process.

| Variables          | Weight | Dry Season | Rainy Season |
|--------------------|--------|------------|--------------|
| Turbidity          | 0.3333 | 2.1458     | 5.6785       |
| Organic Substance  | 0.3333 | 0.3708     | 0.9614       |
| Chlorine Residual  | 0.3333 | 0.3676     | 0.2450       |

In the process capability analysis, the criteria for a process can be said to be capable if the capability index is more than 1. In the dry season or rainy season, univariately, the $P_{pk}$ index for the turbidity variable has a value of more than 1. This means that the water production process is capable to produce water that has a turbidity level within specification limits. However, the $P_{pk}$ index for the organic substance and chlorine residual has a value of less than 1. This indicates that the water production process is not capable of producing water with organic substances and chlorine residual that meets the specification limits.

Based on the $MP_{pk}$ value, the multivariate process of water production during the dry season is incapable of producing water according to the specifications limit. The water production process during the rainy season is capable of producing water within the specification limits. However, it is better to reconsider the weighting for each quality characteristic because univariately the $P_{pk}$ index for the variable organic substance and chlorine residual has a value far from 1 which indicates that the water production process based on the organic substance and residual chlorine is not capable to meet the specification limit.

4. Conclusion

Based on the MEWMC control chart, the variability of the water production process during the dry season and rainy season has been statistically controlled with an optimum $\lambda = 0.1$. Based on the MEWMA control chart, the process mean of water production during the dry season and rainy season with an optimum $\lambda = 0.1$ shows that the process mean for the dry season and the rainy season is not statistically controlled. The main variables that cause the process mean are not statistically controlled in the dry season are organic substance and residual chlorine. The main variable causing the process mean in the rainy season is not statistically controlled as the organic substance. The result of the process capability analysis shows that the water production process is capable of producing water that has a turbidity level within the specification limit, but is not capable of producing water with organic substance and chlorine residual that meet the specification limit. For future research, Trace $R^2$ [12] can be applied in the MEWMC control chart to monitor the variability of the process with a different measurement scale. Also, the Fast Minimum Covariance Matrix (MCD) [13] can be considered in estimating the robust mean vector and covariance matrix.

References

[1] Montgomery D C 2013 *Introduction to Statistical Quality Control (7th ed)* (United States of America: John Wiley & Sons, Inc.)

[2] Johnson R A and D W Wichern 2007 *Applied Multivariate Statistical Analysis (6th ed)* (New Jersey: Prentice-Hall)

[3] Alva J A V and Estrada E G 2009 A Generalization of Shapiro–Wilk’s Test for Multivariate Normality *Communication in Statistics - Theory and Methods* 38 1870-1883

[4] Morrison D F 1967 *Multivariate Statistical Methods (3rd ed)* (USA: McGRAW-Hill Book Company)
[5] Hawkins D M and Maboudou-Tchao E M 2008 Multivariate Exponentially Weighted Moving Covariance Matrix *Technometrics* **50** 155-166

[6] Bodden K M and Rigdon S E 1999 A Program for Approximating the In-Control ARL for the MEWMA Chart *Journal of Quality Technology* **31** 120-123

[7] Vining G and Kowalski S M 2011 *Statistical Methods for Engineers (3rd ed)* (Boston: Cengage Learning)

[8] Kumar D 2006 *Six Sigma Best Practices : A Guide to Business Process Excellence for Diverse Industries* (USA: J. Ross Publishing)

[9] Šibalija T V and Majstorović V D 2010 Process Performance Analysis for Non-Normal Data Distribution *International Journal Total Quality Management & Excellence* **38** 1-4

[10] Raissi S 2009 Multivariate Process Capability Indices on The Presence of Priority for Quality Characteristics *Journal of Industrial Engineering, International* **5** 27-36

[11] Lowry C A, Woodall W H, Champ C W and Rigdon S E 1992 A Multivariate Exponentially Weighted Moving Average Control Chart *Technometrics* **34** 46-53

[12] 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 Engineering* **6** 1665949

[13] Ahsan M, Mashuri M, Lee M H, Kuswanto H and Prastyo D D 2020 Robust Adaptive Multivariate Hotelling’s $T^2$ Control Chart based on Kernel Density Estimation for Intrusion Detection System *Expert Systems with Applications* **145** 113105