PARAMETER
JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA

e-ISSN. 2829-6303
Vol. 04 No. 2, August 2025
Page 197-208

Clustering of Water Quality Location Using Self Organizing Maps (SOM)

Rahmatin Nur Amalia[1], M. Difa Farady[2], Diaz Fitra Aksioma[3], Muhammad Ahsan[4*]

[1,2,3,4]Department of Statistics, Faculty of Science and Data Analytics,
Institut Teknologi Sepuluh Nopember (ITS)
Raya ITS Sukolilo St, Surabaya, 60111, Jawa Timur, Indonesia

E-mail Correspondence Author: *muh.ahsan@its.ac.id*

Abstract

A decline in the number of locations meeting drinking water quality standards was observed based on internal monitoring in 2021 and 2022. To address this, clustering was performed on water quality test locations using Self Organizing Maps (SOM). The analysis of data from 60 locations, considering turbidity, pH, iron, and nitrite parameters, indicated very good water quality. Outliers were detected before clustering, with the Ireng location being the most extreme, showing turbidity of 4.95 NTU and pH of 8.41, near specification limits. The clustering process removed one outlier, forming two clusters with a silhouette coefficient of 0.668. Multivariate normality tests showed the samples were not multivariate normal, leading to use of The Kruskal-Wallis Test. The results revealed significant differences between clusters 1 and 2, particularly in turbidity and iron levels. Cluster 2 had better water quality, with lower turbidity and iron content. Some locations in cluster 1 exceeded 1 NTU turbidity and had higher iron levels. Therefore, the company should improve water quality monitoring and control at locations approaching specification limits.

Keywords: Clustering, drinking water, quality, self-organizing maps.

: https://doi.org/10.30598/parameterv4i1pp197-208

1. INTRODUCTION

Water is a vital component in daily life and the fulfillment of human needs. The quality of water used for human consumption must meet the established standards to ensure public safety and health [1]. To maintain and improve the quality of its products, evaluating the quality of water received by consumers is necessary. The Indonesian government, through Regional Water Utility Companies (PDAM), plays a key role in managing raw water to produce clean water that meets drinking water quality standards. PDAM is not only responsible for water quality but also plays a significant role in ensuring the quantity and sustainability of clean water supply to the public, thereby improving their life quality [2].

The water treatment plant has been conducting customer satisfaction surveys periodically since 2012 to measure the level of customer satisfaction with the services provided by the company. The customer satisfaction survey result in 2021 was 88.92%, and in 2022 was 91.29%. This indicates that there are still customers who are dissatisfied with the services provided by the company. In addition to customer satisfaction surveys, the company also conducts internal monitoring to test several locations to determine whether they meet the drinking water quality standards (MSAM). In 2021, 98.85% of sample points met the MSAM standard, and in 2022, 96.88% of sample points met the MSAM standard. There was a decrease in the percentage of sample points meeting the MSAM standard from 2021 to 2022. Monitoring the water quality can be carried out using control charts [3], [4], [5], [6]. Another point of view is that we are clustering data based on location.

Cluster analysis based on testing locations can be an effective method to determine the extent of the distribution of water that meets established health standards. This method aims to form groups where objects within a group share many similarities, while also significantly different from objects in other group [7], [8]. A popular method in cluster analysis is The K-Means algorithm, which uses the average of all points in a cluster as cluster center (centroids) [9]. In other hand, The K-Means algorithm becomes less effective if the data contains outliers [10], [11]. Moreover, there are some assumptions that must be met, such as representativeness of sample as well as absence of multicollinearity [11].

Self-Organizing Map (SOM) is one of the most renowned clustering algorithms and serves as an effective visualization tool [12], [13], [14], [15]. The SOM method can be considered a spatial data-focused version of The K-Means algorithm. Analogously, each unit represents a group, and the number of groups is determined by the grid dimensions, which are typically arranged in a square or hexagonal shape [16]. The SOM algorithm does not require assumptions since it is a neural network algorithm [17]. The application of the Self-Organizing Maps method in this study is expected to provide a better understanding of the variations in water quality across different testing locations [18], [19]. The results of this cluster analysis can serve as a foundation to optimize water quality management strategies, enhance monitoring efficiency, and provide a prompt response to changes in water quality at various water distribution locations.

2. RESEARCH METHODOLOGY

2.1. Data Sources

This study utilized secondary data obtained from the laboratory water quality tests conducted during the first period of 2023 (January to March). The water quality tests were carried out at 60 water distribution sites. The selection of testing locations was done

randomly, with the chosen locations representing each subdistrict in Mataram City and Lombok Barat Regency.

2.2. Research Variables

The research variable used in this study included five water quality parameters. **Table 1** described those variables along with their specific boundaries.

Table 1. Research Variable

Variable	Unit	Specification
Turbidity (X_1)	NTU	≤ 5
pH (X_2)		6.5 – 8.5
Iron Concentration (X_3)	mg/L	≤ 0.3
Nitrite Concentration (X_4)	mg/L	≤ 3
Temperature (X_5)	Celsius	3 Deviation

2.3. Data Structure

The data structure used in this study is described in **Table 2.** In this structure, $\boldsymbol{X_{i,j}}$ mean a measurement of $\boldsymbol{i}$-th variable on $\boldsymbol{j}$-th location.

Table 2. Data Structure

Sample Number	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$
1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$	$X_{5,1}$
2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$	$X_{5,2}$
3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{4,3}$	$X_{5,3}$
⋮	⋮	⋮	⋮	⋮	⋮
60	$X_{1,60}$	$X_{2,60}$	$X_{3,60}$	$X_{4,60}$	$X_{5,60}$

2.4. Research Stages

The research steps applied in this study are as follows:

1. Describing the characteristics of the data.
2. Conducting clustering using Self-Organizing Maps (SOM).
3. Evaluating the clustering results using the silhouette coefficient.
4. Describing the formed clusters using descriptive statistics.
5. Conducting multivariate normal distribution testing.
6. Analyzing the differences between the formed clusters.
7. Interpreting the clusters.
8. Drawing conclusions

3. RESULT

3.1. Data Characteristics

Before conducting more in-depth analysis, it was important to perform an analysis using descriptive statistics to examine the characteristics of the data used in this study. Descriptive statistics will utilize several measures of central tendency, such as mean, minimum, maximum, standard deviation, and median. Descriptive statistics for each water quality variable will be compared with the specification limits established by the Minister of Health Regulation (Permenkes) No. 492/MENKES/PER/IV/2010 regarding Drinking Water Quality Requirements. The results of the quantitative descriptive statistical calculations as presented in **Table 3**.

The average turbidity was 0.932 NTU, indicated that water was generally clear. The low standard deviation of 0.818 NTU showed that the variation in turbidity across different samples was not enormous. The median turbidity was 0.815 NTU, which was

lower than the mean, indicated that most samples had turbidity below average. All samples met health standards as the maximum value did not exceed the specified limits.

Table 3. Descriptive Statistics

Variable	Mean	StDev	Median	Minimum	Maximum
Turbidity	0.932	0.818	0.815	0.13	4.95
pH	7.25	0.448	7.225	6.53	8.41
Iron	0.009	0.009	0.005	0	0.038
Nitrite	0.06	0.011	0.06	0.04	0.11
Temperature	26.35	2.231	27	22	29

The average pH was 7.25, indicated that the water was neutral to slightly alkaline. The standard deviation of 0.448 showed that there was little variation in pH values across the samples, indicated good consistency. The median pH of 7.225 was around to the mean, suggested a symmetric distribution of pH values. The minimum pH value of 6.53 was within safe range according to Permenkes 2010, which was between 6.5 and 8.5. The maximum pH value of 8.41 approached the upper limit, but still within considerable range.

The iron content in the water was very low with an average of 0.009 mg/L. The small standard deviation of 0.009 mg/L indicated that the variation was small among the samples. The median iron content was 0.005 mg/L, lower than the mean, indicated that most samples had minimal iron content. Based on the minimum and maximum values, all samples were within safe limits.

The average nitrite concentration was 0.06 mg/L, showing high consistency in nitrite measurements. The low standard deviation of 0.011 mg/L indicated small variation among the samples. The median nitrite concentration is same as the mean, indicated a symmetric data distribution. The minimum and maximum nitrite value was 0.04 mg/L, and 0.11 mg/L respectively, were below the upper limit, indicating that all samples had safe nitrite levels.

The average water temperature was 26.35°C, indicating that the water was within a comfortable temperature range. The standard deviation of 2.231°C showed some variation in water temperature among samples. The median temperature was 27°C, slightly higher than the mean, indicated that most samples had a slightly warmer temperature. The minimum and maximum temperature was 22°C and 29°C respectively, with some samples exceeding the maximum allowable temperature of 28°C as specified by Permenkes 2010. Monitoring water temperature was important to ensure comfort and optimal water quality.

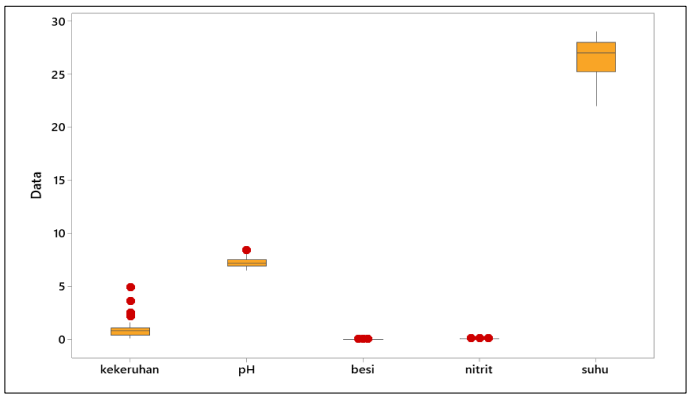

Figure 1. Boxplot of Variables

After conducting descriptive statistics, a descriptive statistical visualization performed with boxplots for the five variables. Based on **Figure 1**, some outliers are detected in turbidity, pH, iron concentration, and nitrite concentration. However, these outliers still fell within safe limits and did not exceed specification limits. It was also observed that these variables had different scales, for example temperature had significant range than iron and nitrite concentration, which are much smaller. Therefore, standardization of those variables was necessary.

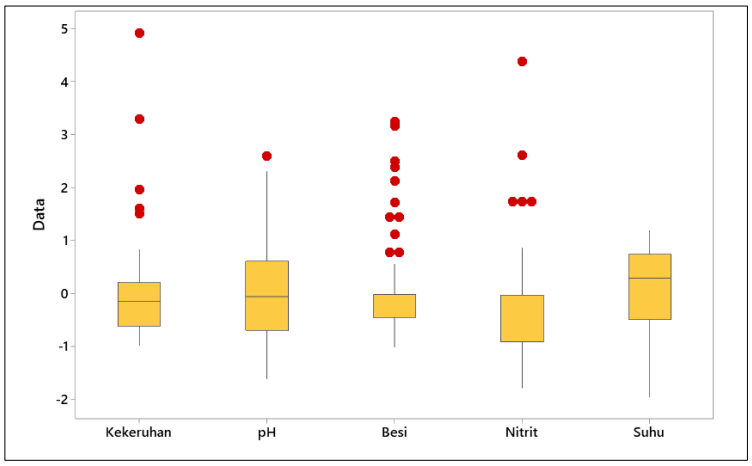

Figure 2. Boxplot of Standardized Variables

Figure 2 showed the result of boxplot visualization after data standardization. It could see that each variable was now on the same scale. This standardization ensures that each variable contributes equally during clustering. Moreover, it helps reduce the impact of dominance from certain variables that previously had larger scales.

3.2. Clustering Results

At this stage, the clustering process using Self Organizing Maps (SOM) was employed **[20], [21]**. It was important to note that this clustering process used the standardized data, as previously explained. Before proceeding with clustering, it should be noted that the data still contained outliers, which were data points that are significantly different from others in the dataset. The presence of outliers could influence the clustering results because SOM algorithms can be affected by extreme data.

Therefore, this stage would attempt to perform clustering without removing outliers first. This is being done to understand the extent to which outliers affected the final clustering results. The SOM parameters used in this study were in default settings from the R function, with an initial learning rate of 0.05.

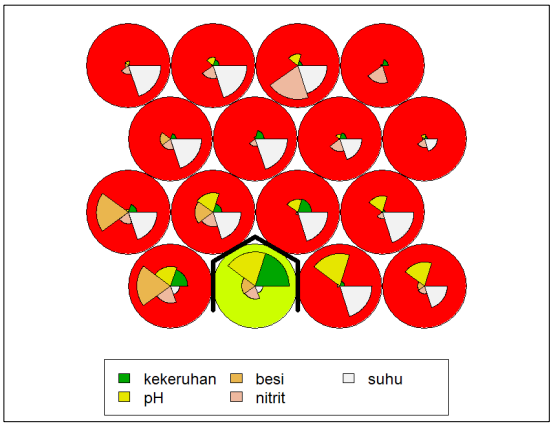

Figure 3. Clustering Without Removing Outliers (2 Clusters)

Based on **Figure 3,** two clusters are formed, with cluster 1 marked in red and cluster 2 in yellow. The temperature variable played a dominant role in the formation of cluster 1, indicated by the large white fan. Cluster 2 consisted of only one member, which was Ireng Sites in Jatisela Village. That location was the most extreme outlier, with turbidity was 4.95 NTU and pH was 8.41. Although these values were still below the specification limits set by Permenkes 2010, they were very close to the specification limits.

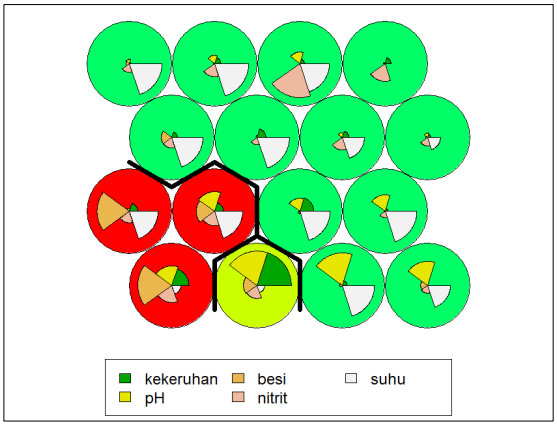

Figure 4. Clustering without Removing Outliers (3 Clusters)

Based on **Figure 4**, three clusters are formed, with cluster 1 in red, cluster 2 in yellow, and cluster 3 in green. There was still a cluster with only one member, which was the most extreme outlier location (Ireng Sites). After performing three clusters without removing outliers, the result showed that outliers still form their own group. This means that outliers had a significant impact on clustering process and could change overall cluster structure. Therefore, to better understand the impact of outliers one approach is to remove them and see the results of clustering performed.

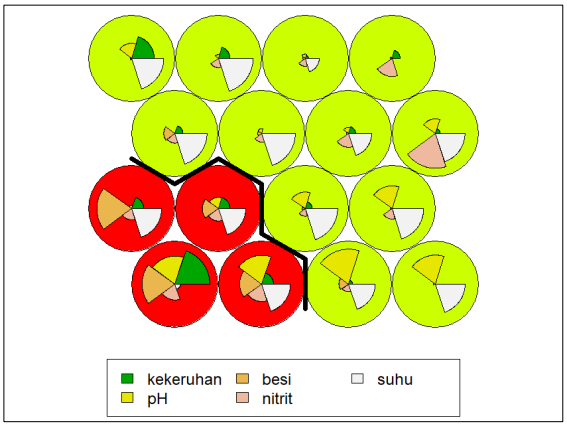

Figure 5. Clustering with Removing Outliers (2 Clusters)

Based on **Figure 5**, cluster 1 is in red, and cluster 2 is in yellow. Cluster 1 had 9 members, with iron concentration variable had significant contribution to formation of this cluster. Cluster 1 included locations with more extreme water quality parameter values compared to Cluster 2. That also showed that in Cluster 1, there are locations with relatively high turbidity, as indicated by the large green fan.

After removing the most extreme outlier, either **Figure 5** or **Figure 6** showed that there are no longer any clusters with just one member.

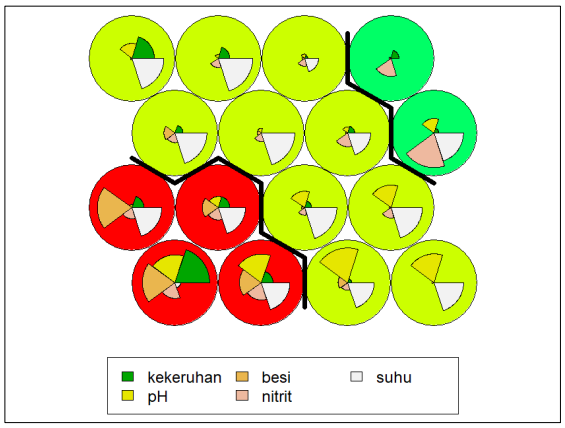

Figure 6. Clustering with Removing Outliers (3 Clusters)

To determine which clustering result is the best, the silhouette coefficient used to evaluate clustering quality **[22], [23]**. Based on **Table 4**, the highest silhouette coefficient value was found in clustering with the removal of one outlier and two clusters, with a value of 0.668. This value indicated that the resulting clusters were of good quality and were well-defined. In clustering evaluation, a silhouette coefficient of 0.668 is categorized as an adequate cluster. Therefore, removing the outlier seemed to have an impact in improving clustering quality.

Table 4. Silhouette Coefficient Evaluation of Clustering

Options	Cluster Number	Silhouette Coefficient	Interpretation
Without removing outlier	2	0.659	Adequate cluster
	3	0.474	Weak cluster
Removing one outlier	2	0.668	Adequate cluster
	3	0.487	Weak cluster

3.3. Cluster Difference Analysis

This analysis is carried out using descriptive statistics and hypothesis testing. Descriptive statistics used to provide an overview of the characteristics of each cluster. **Table 5** showed the characteristics of each cluster.

Table 5. Descriptive Statistics in Each Cluster

Cluster	Variable	Mean	StDev	Median	Min	Max
1	Turbidity	1.57	0.923	1.15	0.94	3.63
	pH	7.31	0.566	7.275	6.53	7.95
	Iron	0.028	0.007	0.028	0.019	0.038
	Nitrite	0.064	0.01	0.06	0.05	0.08
	Temperature	25.778	2.539	27	22	28
2	Turbidity	0.737	0.472	0.73	0.13	2.24
	pH	7.213	0.406	7.195	6.55	8.28
	Iron	0.006	0.003	0.005	0	0.016
	Nitrite	0.059	0.012	0.06	0.04	0.11
	Temperature	26.52	2.15	27	22	29

Based on **Table 5**, it could be seen that Cluster 2 tends to have lower turbidity, pH, iron, and nitrite concentration values. In terms of temperature, Cluster 2 was slightly higher than Cluster 1. Overall, Cluster 2 tends to have better water quality compared to Cluster 1.

After descriptive statistical analysis for each cluster performed, hypothesis testing is conducted with following hypothesis statement:

- H_0: The sample follows a multivariate normal distribution
- H_1: The sample does not follow a multivariate normal distribution

The Kruskal-Wallis Test for multivariate normality applied at 5% significance level, resulted a test statistic of W^*= 0.829 and a p-value of 9.292×10^{-7}. Therefore, with a 95% confidence level H_0 is rejected because the p-value was less than 5% significance level. Thus, it concluded that the sample did not follow a multivariate normal distribution.

3.4. Cluster Interpretation

The results of cluster analysis can be visualized using a map. It should be noted that the coordinates used in the map were from the village/sub-district locations of the sites.

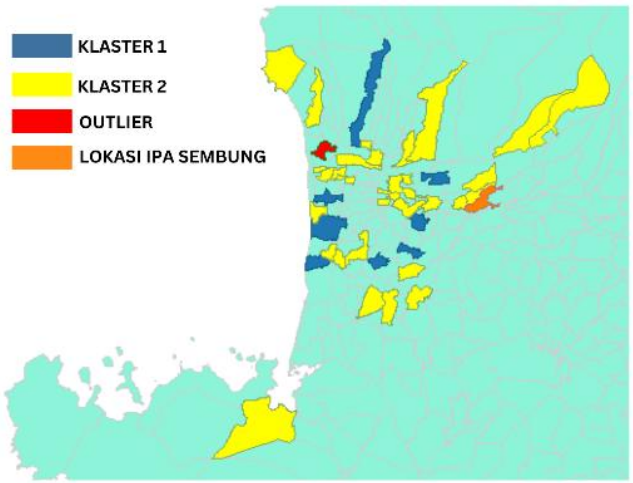

Figure 6. Map of Cluster Results

Based on **Figure 7** the distance between test locations and IPA Sembung seemed to have less effect, as locations relatively close to IPA Sembung could belong to either Cluster 1 or Cluster 2. The factors that might cause variation of water quality at each location could be the physical condition of water pipes or other factors.

Water quality in Cluster 2 had better clarity compared to Cluster 1. However, both still meet the standards set by Permenkes. Locations in Cluster 2 generally had clearer water compared to Cluster 1, although both were still within safe limits. The pH variable of Cluster 1 consisted of several locations with greater variation in pH compared to Cluster 2. Both Cluster 1 and Cluster 2 had some locations with pH levels close to the lower and upper limits.

In addition, for nitrite and iron concentrations, Cluster 2 tends to have lower values compared to Cluster 1. Both clusters had nitrite and iron concentrations that were still far below the specification limits, meaning they were safe. However, compared to Cluster 1, Cluster 2 had more desirable nitrite and iron concentration. For the temperature variable, Cluster 1 included locations with temperature variations that tend to stay within the lower or upper range of the allowed limits. Unlike in Cluster 2, some locations had temperature exceeding the specification limits (reaching 29°C).

From October 2022 to February 2023, pipe repairs were carried out at several points. Cluster 2 consists of locations with better water quality compared to Cluster 1. Interestingly, 11 samples which were part of the pipe repair process are included in Cluster 2. This showed that pipe repairs could improve water quality. Moreover, pipe repairs also reduced the risk of leaks. Pipe leaks not only caused water wastage but also served as entry points for contaminants. With repaired pipes, contamination risks could be minimized, thus ensuring the water supplied was safer to use.

4. CONCLUSION

The analysis of data characteristics shows that turbidity, pH, iron, and nitrite variables in all samples have met the specifications set by Permenkes in 2010. However, there are locations where turbidity and pH values are very close to the specification limits, namely Ireng. Then, for the temperature variable, three locations have temperatures exceeding the specification limit, with a temperature of up to 29°C. The best clustering was achieved by removing one extreme outlier, forming two clusters. The best result is based on the highest silhouette coefficient value obtained, which is 0.668, indicating that the clusters are acceptable. The variables that differ significantly between Cluster 1 and Cluster 2 are turbidity and iron. Cluster 2 tends to have better water quality than cluster 1. The turbidity in Cluster 1 tends to be higher compared to Cluster 2. Some locations in Cluster 1 have turbidity levels exceeding 1 NTU. Similarly, iron content in cluster 1 shows relatively higher levels compared to cluster 2.

It was also important that the best clustering is achieved by removing one outlier, which is Ireng Sites in Jatisela Village. The turbidity level in that site is quite high at 4.95 NTU, which is around specification limit of 5 NTU. In addition, the recorded pH value of 8.41 is also high, but still within allowed specification range (between 6.5 and 8.5).

Therefore, further research and preventive actions at this location are necessary to identify the source of contamination and improve water quality. Steps such as further testing, routine monitoring, and implementing better water management practices may need to be carried out to ensure safe and sustainable water quality for the environment

and local communities. These efforts are expected to help maintain the integrity of water ecosystems and the health of the communities relying on them.

REFERENCES

[1] M. Mashuri, H. Khusna, and F. D. Putri, "Mixed Multivariate EWMA-CUSUM (MEC) Chart based on MLS-SVR Model for Monitoring Drinking Water Quality," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012019.

[2] N. H. D. Asmara, M. Ahsan, M. Mashuri, and H. Khusna, "Quality of Water Production Process Using Mixed Multivariate EWMA-CUSUM (MEC) Control Chart," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012037.

[3] D. D. Prastyo, H. Khusna, M. Mashuri, S. Suhartono, and M. Ahsan, "Multivariate CUSUM control chart based on the residuals of multioutput least squares SVR for monitoring water quality," *MJS*, vol. 38, no. Sp2, pp. 73–83, 2019.

[4] N. Sulistiawanti, M. Ahsan, and H. Khusna, "Multivariate Exponentially Weighted Moving Average (MEWMA) and Multivariate Exponentially Weighted Moving Variance (MEWMV) Chart Based on Residual XGBoost Regression for Monitoring Water Quality.," *Engineering Letters*, vol. 31, no. 3, 2023.

[5] M. Ahsan and T. R. Aulia, "Comparing the Performance of Several Multivariate Control Charts Based on Residual of Multioutput Least Square SVR (MLS-SVR) Model in Monitoring Water Production Process," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012018.

[6] H. Khusna, M. Mashuri, Suhartono, D. D. Prastyo, M. H. Lee, and M. Ahsan, "Residual-based maximum MCUSUM control chart for joint monitoring the mean and variability of multivariate autocorrelated processes," *Prod Manuf Res*, vol. 7, no. 1, pp. 364–394, 2019.

[7] R. A. Johnson and D. W. Wichern, "Applied multivariate statistical analysis," 2002.

[8] M. Y. Matdoan, M. Ahsan, M. Wance, and N. A. Nukuhaly, "Classification of provinces based on the Indonesian Democracy Index using the K-medoids clustering algorithm," in *AIP Conference Proceedings*, AIP Publishing, 2023.

[9] G. Hamerly and C. Elkan, "Learning the k in k-means," *Adv Neural Inf Process Syst*, vol. 16, 2003.

[10] W. I. D. Mining, "Data mining: Concepts and techniques," *Morgan Kaufinann*, vol. 10, no. 559–569, p. 4, 2006.

[11] T. Apriliana and E. Widodo, "Analisis cluster hierarki untuk pengelompokan provinsi di Indonesia berdasarkan jumlah base transceiver station dan kekuatan sinyal," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 3, no. 2, pp. 286–296, 2023.

[12] S. Kusumadewi and H. Purnomo, "Aplikasi Logika Fuzzy untuk pendukung keputusan," *Yogyakarta: Graha Ilmu*, vol. 2, 2010.

[13] F. Bowen and J. Siegler, "Self-organizing maps: A novel approach to identify and map business clusters," *Journal of Management Analytics*, vol. 11, no. 2, pp. 228–246, 2024.

[14] Y. Gajjar, N. Arora, and N. R. Sahoo, "Self-organizing maps: Concept, architecture, and use cases in engineering and finance," in *Deep Learning in Engineering, Energy and Finance*, CRC Press, 2024, pp. 211–249.

[15] E. L. R. Costa, T. Braga, L. A. Dias, É. L. de Albuquerque, and M. A. C. Fernandes, "Self-organizing maps applied to the analysis and identification of characteristics related to air quality monitoring stations and its pollutants," *Neural Comput Appl*, vol. 36, no. 19, pp. 11643–11657, 2024.

[16] R. Wehrens and L. M. C. Buydens, "Self-and super-organizing maps in R: the Kohonen package," *J Stat Softw*, vol. 21, pp. 1–19, 2007.

[17] J. J. Siang, "Jaringan syaraf tiruan dan pemrogramannya menggunakan Matlab," *Penerbit Andi, Yogyakarta*, vol. 11, 2005.

[18] T. Li, G. Sun, C. Yang, K. Liang, S. Ma, and L. Huang, "Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes," *Science of the Total Environment*, vol. 628, pp. 1446–1459, 2018.

[19] Q. Gu *et al.*, "Characterizing the spatial variations of the relationship between land use and surface water quality using self-organizing map approach," *Ecol Indic*, vol. 102, pp. 633–643, 2019.

[20] L. V Fausett, *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India, 2006.

[21] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans Neural Netw*, vol. 11, no. 3, pp. 586–600, 2000.

[22] P. J. Rousseeuw and L. Kaufman, *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons Hoboken, New Jersey, 2009.

[23] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, IEEE, 2020, pp. 747–748.