K-Means Clustering for Grouping Rivers in DIY based on Water Quality Parameters

M. Andang Novianta [1,5], Syafrudin Syafrudin [2,4], Budi Warsito [3,4]

[1] *Students Study Program of Doctoral Environmental Science, School of Postgraduate Studies, Diponegoro University, Indonesia*
[2]*Department of Environmental Engineering, Faculty of Engineering, Diponegoro University, Indonesia*
[3]*Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Indonesia*
[4]*Study Program of Doctoral Environmental Science, School of Postgraduate Studies, Diponegoro University, Indonesia*
[5]*Department of Electrical Engineering, Faculty of Industrial Technology, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia*

[1]m_andang@akprind.ac.id

Abstract - The Special Region of Yogyakarta (DIY) has rivers that cross rural and urban areas that are still used by the community and industry. However, cases of river water pollution in DIY are a major issue in 2021. It is very important to classify rivers according to class so that further analysis and action can be carried out. This study conducted a grouping analysis of rivers in DIY based on water quality parameters such as Total Suspended Solid (TSS), Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Phosphate, Fecal Coli, and Total Coliform. The grouping method uses the K-means algorithm. The data source is secondary data from the DIY Provincial Environment and Forestry Service. The data is in the form of 56 river samples observed in November 2020. The description of the data shows that the average of the 56 river water samples is 24.95 for TSS, 8.84 for DO, 4.33 for BOD5, 20.36 for COD, 0.54 for Phosphate, 22,820 for Fecal Coli, and 59,210 for Total Coliform. The results of grouping with k=6 are the best compared to k = 2, 3, 4, 5, 7, and 8. The number of members in this grouping is n1 = 14, n2 = 1, n3 = 1, n4 = 5, n5 = 18, and n6 = 17. The cluster that has the highest average TSS, BOD, and COD values is the 3rd cluster (Rivers in Bantul and Sleman Regencies). The cluster that has the highest DO value is the 6th cluster (Rivers in Bantul Regency). The cluster that has the highest average Phosphate value is the 2nd cluster (Rivers in Bantul, Sleman, and Gunungkidul Regencies). The cluster that has the highest average Fecal Coli and Total Coliform values are the 4th cluster (Rivers in Bantul Regency, Yogyakarta City, and Sleman Regency).

Keywords: K-Means clustering, rivers classifying, water quality.

I. INTRODUCTION

The Special Region of Yogyakarta (DIY) Environment and Forestry Service said that river water pollution is one of 17 DIY environmental issues or problems in 2021. In addition, it is also one of the three main issues that are a priority in improving environmental quality in DIY with the issue of waste and land conversion that are not in accordance with spatial planning.

Water resources are natural resources that are very important to support the needs of all living things. Water is used in various aspects of life, such as household activities, drinking needs, and other activities. Water resources is divided into two, namely surface water and groundwater. Surface water that is often used by humans is river water, while groundwater that is often used by humans is well water. All of these water sources must always be maintained so that living things can live and reproduce.

DIY has several rivers that flow through urban and rural areas. Many things affect the quality of river water, including the population growth, human activities, and industry. The rate of population growth has led to an increase in settlements in river basins. This makes controlling river water quality more difficult. This also has an impact on the management of domestic waste in river water which is not yet optimal. The Central Statistics Agency states that the average population growth rate in DIY in 2020 is 1.01%. The highest population density in the city of Yogyakarta is 13,413 people/km2 [1].

Based on calculations, 10 rivers have polluted conditions. The parameters of fecal coliform bacteria and

total coliform have a major contribution as sources of contaminants that cause the low value of the pollution index. The high parameter of the coli bacteria indicates that domestic waste management has not been handled properly.

Given the role of river water quality in protecting ecosystems and human life, it is necessary to analyze river water quality. Each river flow in DIY has different quality and pollutant characteristics. Therefore, it is necessary to carry out a location grouping analysis to obtain information on which locations have a high potential for experiencing water pollution. Reference [2] have conducted an analysis of the evaluation of river water quality using the hierarchical clustering method. This study also states that it is important to classify rivers according to their class so that further analysis and action can be carried out. Reference [3] has also used the clustering method which is useful for obtaining water quality ratings, classifying water quality distribution characteristics, knowing variations in pollutant characteristics at each location and time, and finding short-term pollutant conditions.

Many grouping methods can be used, such as Fuzzy C-Means [4], Multi-Layer Perceptron [5], ANFIS [6]-[7], Naive Bayes [8], and others. In this study, researchers used one of them, namely clustering with the K-means algorithm. The K-means method is an unsupervised machine-learning method for grouping observations based on defined characteristics. K-means is a data clustering method for partitioning existing data into one or more groups so that data with the same characteristics can be grouped into the same group [9].

K-means is included in cluster analysis where k is the number of clusters. According to [10], the K-means algorithm is a algorithm to run and implement, because K-means has the ability to group large amounts of data with relatively fast and efficient computational time and is adaptable. The concept in k-means is to get the minimum variation value where each cluster with the distance between the data and the center point of the cluster must be minimum. If in a cluster there are still relatively large variations, the cluster can still be split into two different clusters.

To determine the optimal number of clusters, researchers can use various methods such as the Silhoutte method, the Elbow method, or with a predetermined number of clusters. To get the optimal number of clusters, the Elbow method is used. This method is a method used to determine the best number of clusters by looking at the percentage of the results of the comparison between the number of clusters that form an angle at a point.

Several studies using the K-means method are [11] to identify homogeneous areas of groundwater quality. Reference [12] used the K-means method to classify the status of water quality in rivers in Banjarmasin, Indonesia. Then [13] grouped Balinese handicraft products using the K-Medoids algorithm. Research conducted by [14] to analyze cyberbullying through Instagram and [15] is used for classifying store sales data. Then [16] identified the availability of health human resources in Central Java. In addition, [17] conducted research on the performance of PDAMs in providing water quality, namely based on healthy and unhealthy, unhealthy and sick, and healthy and sick features. In [18], the K-Means algorithm is used to classify poverty in Papua. Reference [19] also uses the k-medoids method.

Based on the problems discussed, this research applies the K-means clustering method to classify rivers in DIY based on water quality parameters. The research results are expected to form river groups. Furthermore, the research results also provide information on rivers that have the same characteristics based on the water quality parameters used and the potential of existing pollutants.

II. METHOD

The source of the data in this study was secondary data from the Yogyakarta Special Region (DIY) Environment and Forestry Service in the 2020 DIY Environmental Quality Index book [20]. The data is in the form of 56 river samples observed in November 2020. Variables include Total Suspended Solid (TSS), Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Phosphate, Fecal Coli, and Total Coliform.

The research steps in using the K-means clustering method are as follows.

1. Prepare river water quality data
2. Performing assumption tests for K-means clustering. This test includes multicollinearity test and outlier detection test. The multicollinearity test is carried out by looking at the output correlation value between variables where the correlation value is not more than 0.95. Meanwhile the detection of outlier data uses a boxplot.
3. Before conducting an analysis with k-means, first, standardize the data if the variables used have different units. Standardization aims to standardize data values that have an inconsistent input format between one variable and another. It use formula in (1).

$$z_i = \frac{x_i - \bar{x}}{\sigma} \quad (1)$$

where x_i is variable data i, $\bar{x}$ is average, and σ is standard deviation

4. Perform the K-means clustering algorithm:
   a. Determine the number of clusters (k) with the Elbow method. This study uses k = 2, 3, 4, 5, 6, 7, and 8
   b. Determine the centroid value or cluster center point
   c. Calculate the distance of each centroid point to the point of each object
   d. Grouping data into clusters with the closest distance
   e. Calculating the new cluster center by finding the average value of the data that is a member of the cluster
   f. Repeating steps 'b' to 'e' so that no data is moved to another cluster.
5. Comparing the clustering results at k = 2, 3, 4, 5, 6, 7, and 8 using the standard deviation ratio, the value of the F test statistic, and the value of the silhouette index coefficient.
6. Conduct cluster member profiling based on the average value

The steps in K-means analysis are assumption test, cluster formation, and cluster validation and profiling. Fig. 1 shows the steps in forming a cluster.

III. RESULT AND DISCUSSION

A. Data Description

Descriptive analysis can be used to describe the characteristics of river water quality data in 56 samples based on specialist doctors, namely, Dissolved Oxygen

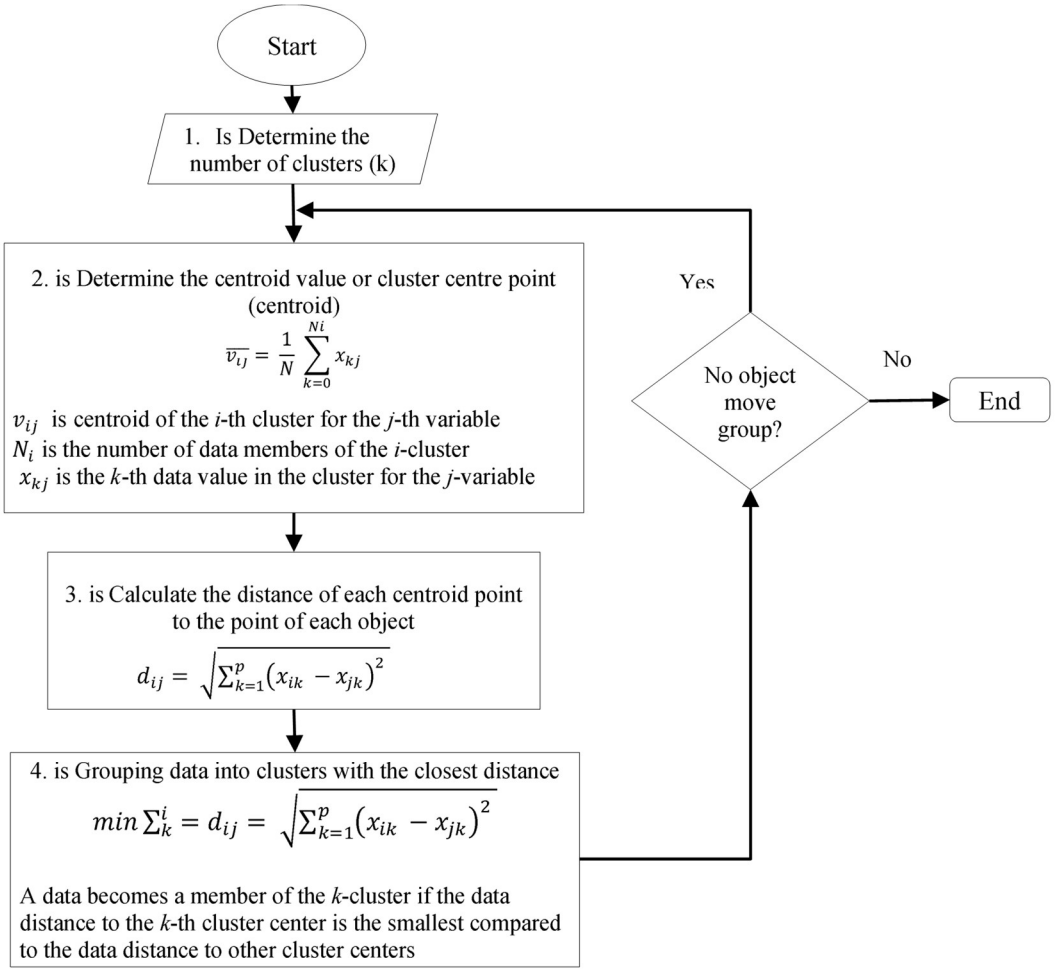

Fig 1. Algorithm of K-means

(DO), Total Suspended Solid (TSS), Biological Oxygen Demand (BOD), Phosphate, Chemical Oxygen Demand (COD), Total Coliform, and Fecal Coli. This descriptive analysis is presented in summary form as shown in Table I which includes the highest score (maximum), lowest score (minimum), average (mean), and standard deviation.

The TSS variable has a minimum value of 0.02, a maximum value is 147.00, an average value is 24.95, and a standard deviation is 29.38. The quality standard is 50 and the number of samples exceeding the quality standard is 5 samples. The average value for the DO variable is 8.84, the minimum value is 4.42, the maximum value is 63.10, and the standard deviation value is 7.52. The quality standard is 4 and the number of samples that exceed the quality standard is 23 samples.

The average value for the BOD5 variable is 4.33, the minimum value is 0.30, the maximum value is 21.30, and the standard deviation value is 2.94. The quality standard is 3 and the number of samples that exceed the quality standard is 21 samples. The average or mean value for the COD variable is 20.36, the lowest or minimum value is 1.40, the maximum value is 72.00, and the standard deviation value is 11.43. The quality standard is 25 and the number of samples that exceed the quality standard is 3 samples.

The average or mean value for the Total Coliform variable is 59,210, the lowest or minimum value is 90.00, the highest or maximum value is 920,000, and the standard deviation value is 140,598. The quality standard is 25 and the number of samples exceeding the quality standard is 9 samples (Table II).

TABLE I
DESCRIPTIVE STATISTICS

Variable	Minimum	Average (Mean)	Maximum	Standard deviation
TSS	0.02	24.95	147.00	29.38
DO	4.42	8.84	63.10	7.52
BOD5	0.30	4.33	21.30	2.94
COD	1.40	20.36	72.00	11.43
Fosfat	0.07	0.54	1.48	0.33
Fecal Coli	90	22,820	350,000	54,873
Total Coliform	90	59,210	920,000	140,598

TABLE II
NUMBER OF SAMPLES BASED ON WATER QUALITY STATUS

Water Quality Status	Frequency
Fulfillment	3
Mild	31
Moderate	18
Severe	4
Total	**56**

Table II shows the number of samples based on water quality status. Fulfillment status has a frequency of 3. Mild status has a frequency of 31. Moderate status has a frequency of 18. Severe status has a frequency of 4. The total water quality status is 56.

B. K-means Clustering Algorithm

In using K-means clustering, the first step is to determine the optimal number of clusters using the Elbow method. This method is one of the methods that is often used to determine the optimal number of clusters by looking at the percentage of the results of the comparison between the number of clusters that will form an angle at a point. The results of the Elbow method are presented in Fig. 2. It can be seen that the line that has a fracture that forms an elbow is at k = 6, meaning that using the Elbow method the best number of clusters is obtained, namely 6 clusters. However, to see the comparison of the number of clusters, calculations are performed using k = 2, k = 3, k = 4, and k = 5, k = 6, k = 7, and k = 8.

After getting the optimal number of clusters, then do clustering with k-means algorithm. The results of the number of members in each cluster are presented in Table III.

Fig. 3 shows the visualization of the grouping results in each cluster (k). This visualization is formed from two dimensions. Dimension 1 explains the clustering result of 31.5% and dimension 2 explains 24%. Good clustering is indicated by high homogeneity between observations within the cluster and high heterogeneity between clusters. High homogeneity between observations within the cluster is shown by the locations of the observations that are close together. Meanwhile, high heterogeneity between clusters is indicated by the large distance between clusters.

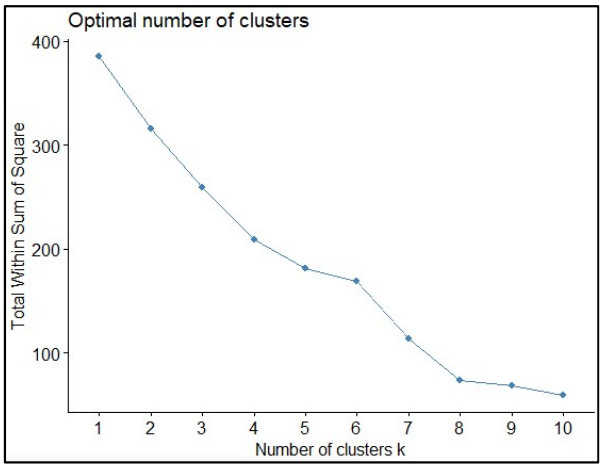

Fig. 2 Elbow method graph

TABLE III
NUMBER OF MEMBERS IN EACH CLUSTER k = 2, 3, 4, 5, 6, 7, and 8

Numbers of Cluster (k)	Number of cluster members	
2	Cluster 1 = 35	
	Cluster 2 = 31	
3	Cluster 1 = 16	
	Cluster 2 = 6	
	Cluster 3= 34	
4	Cluster 1 = 19	Cluster 3= 25
	Cluster 2 = 1	Cluster 4= 11
5	Cluster 1 = 15	Cluster 4= 33
	Cluster 2 = 1	Cluster 5= 6
	Cluster 3= 1	
6	Cluster 1 = 14	Cluster 4= 5
	Cluster 2 = 1	Cluster 5= 18
	Cluster 3= 1	Cluster 6= 17
7	Cluster 1 = 14	Cluster 5= 1
	Cluster 2 = 1	Cluster 6= 14
	Cluster 3= 12	Cluster 7= 9
	Cluster 4= 5	
8	Cluster 1 = 12	Cluster 5= 1
	Cluster 2 = 17	Cluster 6=5
	Cluster 3= 1	Cluster 7= 13
	Cluster 4= 6	Cluster 8= 1

If you look at the visualization comparison of grouping, grouping with k = 2, 3, and 7 is better than k = 4, 5, 6, and 8. As an illustration in k = 3, observations in cluster 1 are close together and denoted in red. Likewise, the observations in cluster 2 are close together and denoted in green. Furthermore, the observations in cluster 3 are close together and denoted in blue. These three groups also have large distances or colors that do not overlap.

After knowing the results of clustering, then do a comparison to get the best grouping results. The method used is to compare the value of the standard deviation ratio and the value of the F test statistic from MANOVA. In addition, validation was also carried out to find out whether the cluster results obtained were valid to use or not. The method used is to look at the value of the silhouette index coefficient. The results of the standard deviation ratio, and the value of the F test statistic, and the value of the silhouette index coefficient are presented in Table IV.

TABLE IV
COMPARISON OF STANDARD DEVIATION RATIO, F TEST STATISTICAL VALUE, AND SILHOUETTE INDEX COEFFICIENT VALUE

Number of clusters (k)	Standard Deviation Ratio	F Test MANOVA	*Silhouette index Coefficient Value*
2	4.170	23.340	**0.29**
3	2.534	6.802	0.24
4	1.116	15.453	**0.29**
5	0.637	**29.605**	0.28
6	**0.345**	10.407	0.26
7	0.416	9.6201	0.22
8	0.378	5.1747	0.20

The explanation of each item is as follows:

a. Standard Deviation Ratio
The value of the Standard Deviation Ratio is obtained from the results of dividing the average standard deviation within groups and the standard deviation between groups. A good grouping has a very small average within and has a very large average between, so that the average standard deviation ratio is the smallest. From the comparison results, it can be seen that grouping with k=6 has the smallest standard deviation ratio value of 0.345.

b. F Test MANOVA
The F test on MANOVA uses a hypothesis
Ho: the average BOD, DO, TSS, COD, Phosphate, Fecal Coli, and Total Coliform between clusters are the same
H1: the average BOD, DO, TSS, COD, Phosphate, Fecal Coli, and Total Coliform between clusters is different.
The conclusion is that Ho is rejected if F test statistic value > $F_{5\%,7.48}$ or F > 2.21. If Ho is rejected, this indicates that the average TSS, DO, BOD, COD, Phosphate, Fecal Coli, and Total Coliform are different between clusters. This also means that the observations between clusters already have a large average difference or high heterogeneity between clusters. The value of the F test statistic for all k has exceeded 2.21. This shows that all groupings have shown high heterogeneity between clusters. However, grouping with k=5 has the highest value of 29.605. Therefore, grouping with k=5 can be said to be better than the other k.

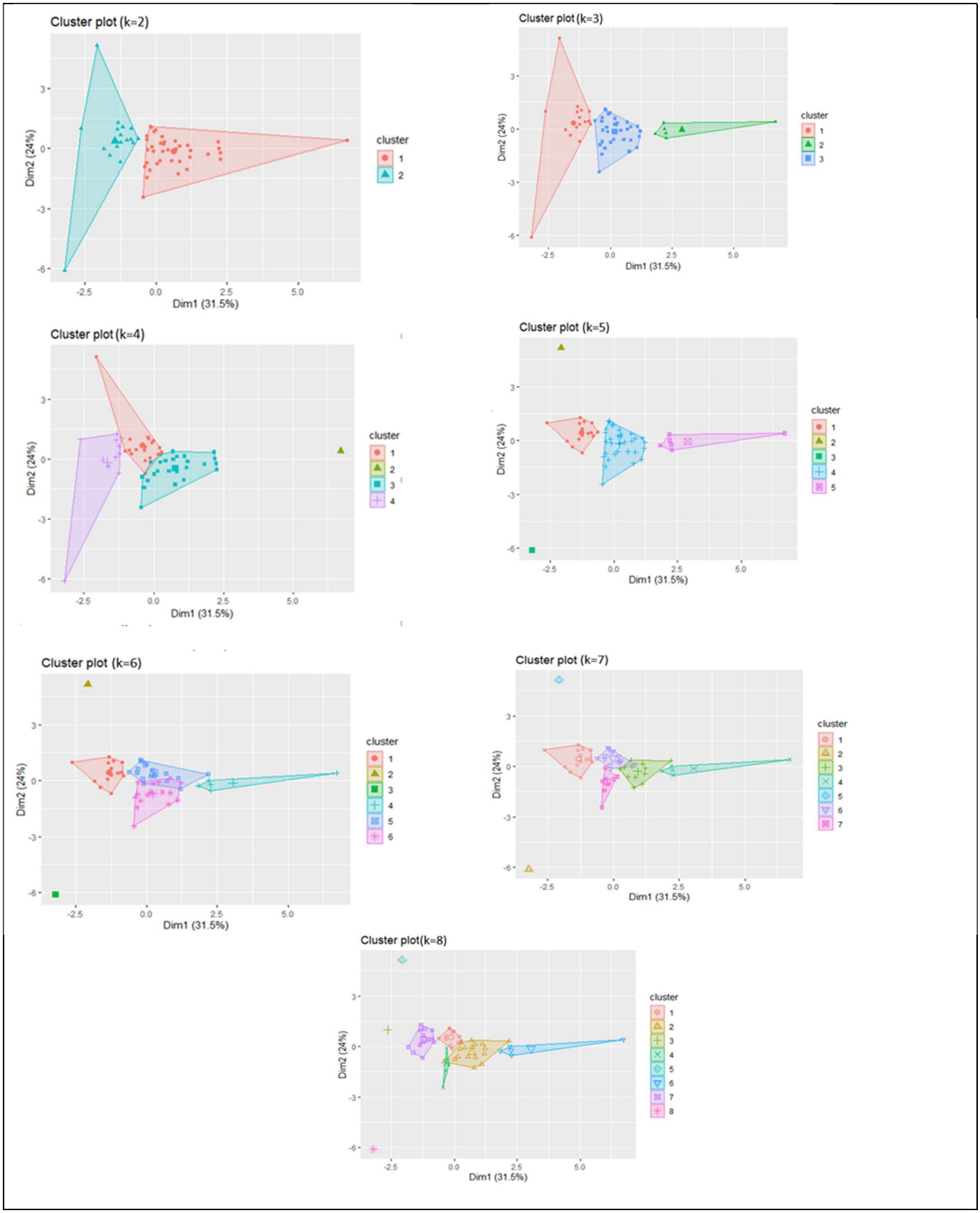

Fig. 3 Grouping result plot with k = 2, 3, 4, 5, 6, 7, 8

c. Silhouette Index Coefficient
The silhouette coefficient method is a combination method between the cohesion and the separation method. The separation method serves to measure how far a cluster is separated from other clusters. The function of the Cohesion method is used to measure how close the relationship is between objects in a cluster.. The silhouette index value that is getting closer to the value 1 then the grouping will be better or valid. Based on the comparison, the greatest silhouette index value is 0.29 with the number of clusters 2 and 4. Visualization of the comparison results is shown in Fig. 4.

This study chooses the results of grouping with k = 6 as the best. Therefore, profiling was carried out at k=6. This profiling aims to know the description and characteristics of each variable in each cluster. Profiling is done based on the average value. Table V shows the profiling of each variable in each cluster with k=6.

The cluster that has the highest average TSS, BOD, and COD values is the 3rd cluster. This cluster consists of the Gajahwong River in Bantul Regency and the Content River in Sleman and Bantul Regencies. The cluster that has the highest average DO value is the 6th cluster, namely the Gajahwong River in Bantul Regency. The cluster that has the highest average Phosphate value is the 2nd cluster. This cluster includes the Winongo River, Gajahwong River, Code River, and Bedog River in Bantul Regency, the Belik River in Sleman Regency, and the Oyo River in Gunungkidul Regency and Bantul Regency. The cluster that has the highest average values of Fecal Coli and Total Coliform is the 4th cluster. This cluster includes the Winongo River and Bedog River in Bantul Regency, the Code River in Yogyakarta City and Bantul Regency, the Kuning River, the Belik River, and the Bulus River in Sleman Regency, as well as the Gajahwong River in Sleman, Yogyakarta, and Bantul Regency.

IV. CONCLUSION

The results of the data description show that there are still many river locations that have levels above the quality standard, including TSS, DO, BOD, COD, Phosphate, Fecal Coli, and Total Coliform. Through k-means clustering analysis with a value of k = 6, the cluster that has the highest average TSS, BOD, and COD values is the 3rd cluster (namely the Rivers in Bantul and Sleman Regencies). The cluster that has the highest average DO value is the 6th cluster (namely the River in Bantul Regency). The cluster that has the highest average Phosphate value is the 2nd cluster (namely the Rivers in Bantul, Sleman, and Gunungkidul Regencies). The cluster that has the highest average Fecal Coli and Total Coliform values are the 4th cluster (namely Rivers in Bantul Regency, Yogyakarta City, and Sleman Regency). The number of observations of each river is $n_1 = 14$, $n_2 = 1$, $n_3 = 1$, $n_4 = 5$, $n_5 = 18$, and $n_6 = 17$.

ACKNOWLEDGEMENT

We would like to say thank you to Department of Environment and Forestry of the Special Region of Yogyakarta for providing research permits in accessing data; and Institut Sains & Teknologi AKPRIND Yogyakarta for funding this research; and Study Program of Doctoral Environmental Science, School of Postgraduate Studies, Diponegoro University for providing the excellent support and cooperation to produce the work described ini this paper.

TABLE V
CLUSTER PROFILING WITH k = 6

Cluster	TSS	DO	BOD	COD	Fosfat	Fecal Coli	Total Coliform
1 (n=14)	64.05	8.35	4.58	15.59	0.24	1,605	4,673
2 (n=1)	11.07	6.83	3.68	20.41	**0.97**	53,015	145,846
3 (n=1)	**85.00**	8.70	**21.30**	**72.00**	0.24	2,000	9,000
4 (n=5)	3.00	7.02	2.69	18.68	0.22	**350,000**	**920,000**
5 (n=18)	8.26	8.10	4.08	21.73	0.52	8,237	16,292
6 (n=17)	54.00	**63.10**	0.30	1.40	0.19	150	1,800

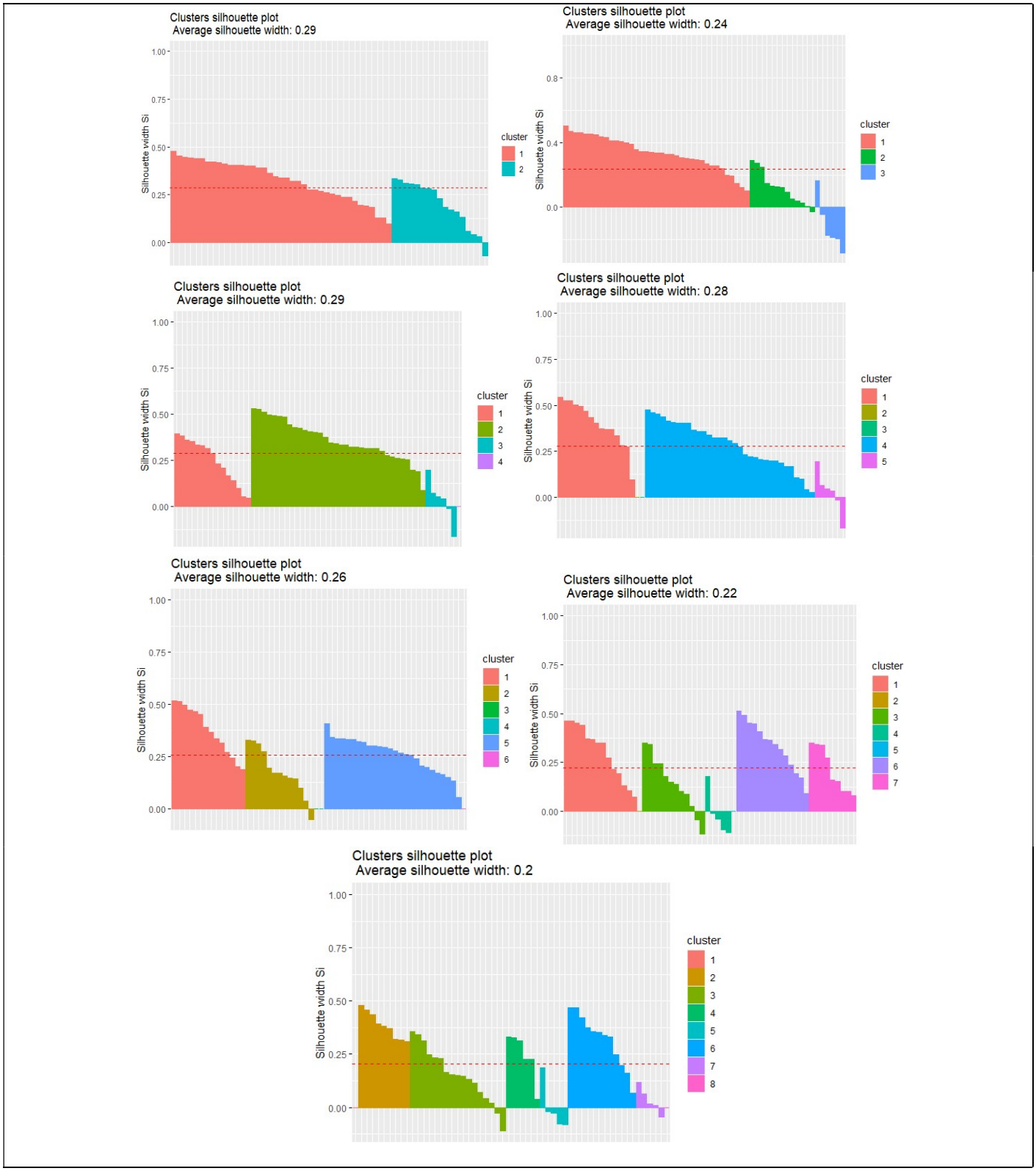

Fig. 4 Silhouette clusters plot with k = 2, 3, 4, 5, 6, 7, and 8

REFERENCES

[1] B. Jogja, "Data Laju Pertumbuhan Penduduk di DIY," 2020. http://bappeda.jogjaprov.go.id/dataku/data_dasar/index/701-penduduk (accessed Jan. 02, 2023).

[2] B. Warsito, S. Sumiyati, H. Yasin, and H. Faridah, "Evaluation of river water quality by using hierarchical clustering analysis," 2021, doi: 10.1088/1755-1315/896/1/012072

[3] Z. Di, M. Chang, and P. Guo, "Water Quality Evaluation of the Yangtze River in China Using Machine Learning Techniques and Data Monitoring on Different Time Scales," *Water*, vol. 11, no. 2, p. 339, 2019, doi: https://doi.org/10.3390/w11020339

[4] O. Herliana, T. S. Widodo, and I. Soesanti, "Klasifikasi Nomsupervised Citra Thermal Kanker Payudara Berbasis Fuzzy C-MEANS," *Jnteti*, vol. 1, no. 3, pp. 1–5, 2012, doi: 10.22146/jnteti

[5] H. A. Nugroho, D. Hardiyanto, and T. B. Adji, "Nipple detection to identify negative content on digital images," in *Proceeding - 2016 International Seminar on Intelligent Technology and Its Application, ISITIA 2016: Recent Trends in Intelligent Computational Technologies for Sustainable Energy*, 2016, pp. 43–48. doi: 10.1109/ISITIA.2016.7828631

[6] D. Hardiyanto, S. Kristiyana, D. Kurniawan, and D. A. Sartika, "Klasifikasi Motif Citra Batik Yogyakarta Menggunakan Metode Adaptive Neuro Fuzzy Inference System," *Setrum Sist. Kendali-Tenaga-elektronika-telekomunikasi-komputer*, vol. 8, no. 2, p. 229, 2019, doi: 10.36055/setrum.v8i2.6545

[7] D. A. Sartika, H. Arrosida, and D. Hardiyanto, "Implementasi Teknik Klasifier Adaptive Neuro Fuzzy Inference System Untuk Mengklasifikasi Motif Citra Batik Jawa Timur," *Setrum Sist. Kendali-Tenaga-elektronika-telekomunikasi-komputer*, vol. 11, no. 1, pp. 126–134, 2022, doi: 10.36055/setrum.v11i1.14872

[8] I. Rish, "An empirical study of the naive Bayes classifier," *IJCAI 2001 Work Empir Methods Artif Intell*, vol. 3, no. 22, pp. 4863–4869, 2001, doi: 10.1039/b104835j

[9] L. Wenchao, Z. Yong, and X. Shixiong, "A Novel Clustering Algorithm Based on Hierarchical and K-means Clustering," in *Chinese Control Conference*, 2007, pp. 605–609, doi: 10.1109/CHICC.2006.4347538

[10] H. Sulatri and A. I. Gufroni, "Penerapan Data Mining Dalam Pengelompokkan Penderita Thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017, doi: https://doi.org/10.25077/TEKNOSI.v3i2.2017.299-305

[11] O. Mohammadrezapour, O. Kisi, and F. Pourahmad, "F. Fuzzy c-means and K-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality," *Neural Comput. Appl.*, vol. 32, no. 8, pp. 3763–3775, 2020, doi: https://doi.org/10.1007/s00521-018-3768-7

[12] T. Zubaidah, N. Karnaningroem, and A. Slamet, "K-means method for clustering water quality status on the rivers of Banjarmasin, Indonesia," *ARPN J. Eng. Appl. Sci.*, vol. 13, no. 6, 2018. doi:10.31227/osf.io/s9n2u

[13] D. A. I. C. Dewi and D. A. K. Pramita, "nalisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, 2019, doi: 10.31940/matrix.v9i3.1662

[14] A. Muhariya, I. Riadi, and Y. Prayudi, "Cyberbullying Analysis on Instagram Using K-Means Clustering," *JUITA J. Inform.*, vol. 10, no. 2, p. 261, 2022, doi: 10.30595/juita.v10i2.14490

[15] F. Indriyani and E. Irfiani, "Clustering Data Penjualan pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means," *JUITA J. Inform.*, vol. 7, no. 2, p. 109, 2019, doi: 10.30595/juita.v7i2.5529

[16] M. Nishom, S. F. Handayani, and D. Dairoh, "Pillar Algorithm in K-Means Method for Identification Health Human Resources Availability Profile in Central Java," *JUITA J. Inform.*, vol. 9, no. 2, p. 145, 2021, doi: 10.30595/juita.v9i2.9860

[17] T. N. Hidayat, F. A. Purnomo, and Y. Yudhanto, "PDAM Performance Clustering using K-Means," in *1st International Conference on Smart Technology, Applied Informatics, and Engineering (APICS)*, 2022, pp. 148–152, doi: 10.1109/APICS56469.2022.9918683

[18] I. M. Erwandi, "Pengelompokan kemiskinan Kabupaten/Kota di Papua dengan menggunakan metode K-Means," UIN Sunan Ampel Surabaya, 2021.

[19] M. A. Nahdliyah, T. Widiharih, and A. Prahutama, "METODE k-MEDOIDS CLUSTERING DENGAN VALIDASI SILHOUETTE INDEX DAN C-INDEX (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018)," *Gaussian*, vol. 8, no. 2, pp. 161–170, 2019, doi: https://doi.org/10.14710/j.gauss.8.2.161-170

[20] Admin, "Indeks Kualitas Lingkungan Hidup (IKLH) DIY," yogyakarta, 2020. [Online]. Available: https://dlhk.jogjaprov.go.id/indeks-kualitas-lingkungan-hidup-iklh-diy.