Evaluation of river water quality by using hierarchical clustering analysis

B Warsito, S Sumiyati, H Yasin, H Faridah

1Department of Statistics, Faculty of Sciences and Mathematics, Diponegoro University, Semarang Indonesia
2Department of Environmental Engineering, Faculty of Engineering, Diponegoro University, Semarang Indonesia
3Doctoral Program of Environmental Science, School of Postgraduate Studies, Diponegoro University, Semarang Indonesia

Abstract. Assessment of water pollution is a critical study because it can affect humans directly. Likewise, river water is widely used for various daily needs. It is important to group rivers according to their classes so that further analysis and action can be carried out. This article discusses the clustering of rivers in several areas in the southeast part of Central Java Province consisting of 14 sampling stations based on several water quality parameters. The pollutant parameters include TSS, electrical conductivity, pH, BOD, COD, and DO. The method used is Hierarchical clustering in which the object grouping begins with grouping two objects with the closest distance being combined into one cluster, and the process continues until one cluster is formed consisting of all objects. The results show that five clusters are the ideal choice. Except for electrical conductivity, the parameters observed are dominantly the difference between clusters. Through the formation of river clusters based on their water quality, the characteristics of each cluster and cluster members with high similarity can be identified.

1. Introduction
Water quality is one of the important criteria in measuring environmental management indicators. A generic water quality standard that satisfies all water uses is challenging [1]. The water quality index can be used to assess the quality of water bodies and the suitability of their designation. Monitoring river quality through measuring the water quality index is important because it can be used as a fulfillment of quality standards. In general, rivers in Central Java, Indonesia, are in the medium category. In the analysis of aquatic systems, modeling water quality parameters is very important [2]. The process of water planning and management relies heavily on the use of statistical models. The model is used to help identify and evaluate alternative ways to meet various planning and management objectives [3]. The development of an efficient model needs to be done using both spatial and temporal data. Surface water quality models can be useful tools for simulating and predicting the level, distribution, and risk of chemical pollutants in a given water body. Modeling result is an important component in environmental impact assessments and can provide the basis for technical support for environmental management agencies to make informed decisions [4]. One type of modeling which is useful in monitoring river water quality is clustering. The main aim of clustering methods is to identify groups of objects, or clusters,
that are more similar to each other than to other clusters [5]. This technique separates objects into groups with high heterogeneity as well as analyzes the similarity between objects in one group.

Various clustering methods have been developed, one of the popular ones is the hierarchical cluster. It is one of the popular clustering techniques in Machine Learning. In hierarchical clustering, the most similar objects will be grouped. Based on the similarity of these clusters, the most similar clusters are combined. This process is repeated until only one cluster remains. For the time needed, k-means and hierarchical clustering take less time to build the model than other methods [6]. However, this method is considered more suitable than k-Means Clustering, where the number of clusters must be decided at the beginning of the algorithm. Ideally, the number of clusters to have is unknown at the beginning of the algorithm. Based on the background, the purpose of this study is to cluster rivers based on water quality based on several pollutant parameters using a hierarchical cluster technique.

2. Methodology

2.1. Data

The research area includes rivers in Central Java Province where water quality measurements have been carried out in the first monitoring period of 21 January 2020. The data used in this study were collected from fourteen monitoring stations under the river water quality monitoring program by the Department of Environment and Forestry, Central Java Province (Table 1). The locations of these rivers are in the regencies of Sukoharjo, Sragen, Karanganyar, Boyolali, and Surakarta.

| Station     | Code | Region    | TSS (mg/l) | DHL (mg/l) | pH  | BOD (mg/l) | COD (mg/l) | DO (mg/l) |
|-------------|------|-----------|------------|------------|-----|------------|------------|-----------|
| Palur1      | PL1  | Sukoharjo | 13.1       | 498        | 7.81| 3.6        | 18.1       | 4.1       |
| Palur2      | PL2  | Sukoharjo | 43.3       | 582        | 7.78| 15.2       | 71.6       | 3.1       |
| Mungkung1   | MK1  | Karanganyar | 11.5     | 358        | 8.22| 2          | 7.6        | 6.8       |
| Mungkung2   | MK2  | Sragen     | 140        | 249        | 8.11| 3.4        | 14.1       | 6.2       |
| Samin1      | SM1  | Karanganyar | 2.5       | 228        | 8.32| 2          | 8.8        | 6.8       |
| Samin2      | SM2  | Sukoharjo  | 21.5       | 617        | 7.14| 46.4       | 211.6      | 1         |
| Jlantah1    | JL1  | Karanganyar | 20.5     | 275        | 7.93| 2          | 6.8        | 6.8       |
| Jlantah2    | JL2  | Sukoharjo  | 11.8       | 373        | 7.75| 2          | 8.5        | 6.6       |
| Grompol1    | GP1  | Karanganyar | 136      | 271        | 7.74| 5          | 20.2       | 5.5       |
| Grompol2    | GP2  | Sragen     | 130        | 289        | 7.7 | 4.8        | 27.8       | 5.4       |
| Baki1       | BK1  | Boyolali   | 20.5       | 563        | 8.04| 2.1        | 12.2       | 6.5       |
| Baki2       | BK2  | Sukoharjo  | 5.2        | 507        | 8.06| 2          | 4.9        | 6.2       |
| Premulung1  | PM1  | Boyolali   | 23         | 544        | 7.81| 2          | 14.4       | 6.2       |
| Premulung2  | PM2  | Surakarta  | 2.5        | 829        | 7.81| 6.6        | 19.5       | 3.1       |

2.2. Water quality parameters

The water pollutant parameters used as the basis for clustering include and Total Suspended Solids (TSS) in mg/l, Electrical Conductivity (DHL) in µg/l, pH in Sorensen scale, Biological Oxygen Demand (BOD) in mg/l, Chemical Oxygen Demand (COD) in mg/l, and Dissolved Oxygen (DO) in mg/l. Based
on the data received, the measurement of water quality parameters was carried out at the Environmental Testing Center and Laboratory, Department of Environment and Forestry, Central Java Province.

2.3. Hierarchical cluster analysis
Clustering is included in multivariate data analysis. This method group objects into clusters by measuring distances and identifying each cluster [7]. Distance metrics that can be used include Euclidean distance and Manhattan distance. The selection of metrics determines the shape of the resulting cluster. This is because clusters that are close together according to one metric can be distant from each other according to another metric. Objects with high similarity will be grouped in one cluster while objects with low similarity will be in different clusters. In other words, the homogeneity within the same cluster is high while the homogeneity between clusters is low. There are various cluster methods such as K-means clustering, hierarchical clustering, K-mode clustering, Fuzzy algorithms, etc.

Each object is treated as a single cluster and then sequentially combined or agglomerated into cluster pairs until all clusters have been combined into a single cluster containing all objects. In other terms, the concept of the hierarchical method begins by combining the two most similar objects, then the combination of these two objects will join again with one or more other most similar objects. This clustering process will eventually agglomerate into one large cluster that includes all objects. Stages of hierarchical clustering can be described as follows [8]:
1. Assign every object into a separate cluster.
2. In each pair of clusters, calculate the pairwise distance. create a matrix whose elements are the calculated distance values.
3. Find the shortest distance from the pair of clusters.
4. Combine the identified pairs by removing both clusters from the distance matrix.
5. Calculate the distances from the new cluster to all other clusters, then update the distance matrix.
6. Repeat these stages up to the matrix is reduced to only a single element.

In hierarchical clustering, it is not necessary to determine the number of clusters at the beginning of the process. This method is also known as the agglomerative method which is presented in a dendrogram [9]. By using a dendrogram, users can consider a visual representation to get the desired clustering. Dendrograms provide a comfortable method for exploring the multiple possibilities of grouping data [10].

3. Results and discussion
In this research, the clustering analysis was applied to standardized raw data. The use of standardization procedure is to eliminate the influence of different units of measurement. Standardization is required when there is unit variability. In this case, electrical conductivity and pH have different units from other parameters. Standardization tends to increase the effect of characteristics with small variances and reduce the effect of characteristics with large variances. Therefore, the data used for data processing is standardized data. The process of forming clusters at observation stations using the hierarchical method begins with calculating the distance matrix between variables using the Euclidean distance. Table 2 shows the distance matrix between one variable and another. The smaller the Euclidean distance, the more similar the two variables are. Based on the closest distance, the variable with the highest similarity will form a group or cluster. After the distance between variables is calculated by the Euclidean distance, the grouping is carried out in stages. The complete agglomeration process is presented in Table 3. In the first stage, by pay attention to the Coefficients column, one cluster is formed consisting of samples no 3 (MK1) and 8 (JL2) with a distance of 0.078. Because the agglomeration process starts from the two closest objects, then the distance is the closest of all the 14 object distance combinations. Next, attention is directed to the last column, the next stage is number 4. This means that the next clustering is done by looking at stage 4. The same technique is continued until the last stage. The calculation of the coefficients in this agglomeration process is complex, especially involving many objects and continues to grow. The agglomeration process will eventually unite all objects into one cluster. In the process, some clusters are generated with each member, depending on the number of clusters formed.
Table 2. Proximity matrix of squared Euclidean distance.

| Case | PL1   | PL2   | MK1   | MK2   | SM1   | SM2   | JL1  | JL2  | GP1  | GP2  | BK1  | BK2  | PM1  | PM2  |
|------|-------|-------|-------|-------|-------|-------|------|------|------|------|------|------|------|------|
| PL1  | 0.000 | 10.287| 2.683 | 3.149 | 5.357 | 30.019| 4.063| 2.231| 2.216| 15.277| 2.137| 1.834| 2.301| 4.790|
| PL2  | 10.287| 0.000 | 16.859| 15.558| 24.926| 19.642| 13.968| 16.025| 14.572| 26.758| 9.861| 18.316| 19.763| 17.863|
| MK1  | 2.683 | 16.859| 0.000 | 6.111 | 1.248 | 42.999| .949 | .078 | .832 | 15.969| 2.109| 1.156| 1.995| 11.667|
| MK2  | 3.149 | 15.558| .611  | 0.000 | 1.349 | 39.998| .577 | .687 | .248 | 14.876| 3.548| 2.894| 4.059| 14.454|
| SM1  | 5.357 | 24.926| 1.248 | 1.349 | 0.000 | 47.855| 2.938| 1.523| 1.704| 17.267| 6.474| 2.661| 3.357| 15.419|
| SM2  | 30.019| 19.642| 42.999| 39.998| 47.855| 0.000 | 43.690| 41.688| 38.211| 45.077| 37.223| 41.001| 40.277| 29.919|
| JL1  | 4.063 | 13.968| .949  | .577  | 2.938 | 43.690| .000 | .986 | .922 | 15.159| 2.704| 3.855| 5.311| 16.462|
| JL2  | 2.231 | 16.025| .078  | .687  | 1.523 | 41.688| .986 | .000 | .719 | 14.054| 1.864| 1.024| 1.773| 10.849|
| GP1  | 2.216 | 14.572| .832  | .248  | 1.704 | 38.211| .922 | .719 | .000 | 13.300| 3.454| 2.624| 3.629| 12.730|
| GP2  | 15.277| 26.758| 15.969| 14.876| 17.267| 45.077| 15.159| 14.054| 13.300| 17.624| 16.049| 16.904| 16.444| 25.072|
| BK1  | 2.137 | 9.861 | 2.109 | 3.548 | 6.474 | 37.223| 2.704| 1.864| 3.454| 17.624| .000 | 2.206| 2.962| 8.223 |
| BK2  | 1.834 | 18.316| 1.156 | 2.894 | 2.661 | 41.001| 3.855| 1.024| 2.624| 16.904| 2.206| .000 | .164 | 6.202 |
| PM1  | 2.301 | 19.763| 1.995 | 4.059 | 3.357 | 40.277| 5.311| 1.773| 3.629| 16.844| 2.962| .164 | .000 | 5.343 |
| PM2  | 4.790 | 17.863| 11.667| 14.454| 15.419| 29.919| 16.462| 10.849| 12.730| 25.072| 8.223 | 6.202| 5.343| .000 |

Dendrogram is useful for showing the existing cluster members according to the number of clusters that should be formed. Figure 1 presents the dendrograms for the average linkage, for the variables TSS, Electrical Conductivity, pH, BOD, COD, and DO whereas, Table 4 lists the clusters formed by this method.
Figure 1. Dendrogram results of clustering rivers in 14 locations based on water quality parameters.

Table 4. Cluster membership of rivers at the observation locations based on water quality samples.

| Number of Clusters | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
|--------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 6                  | SM2       | GP2       | PL2       | PM2       | SM1, BK2, PM1, PL1, BK1 | MK1, JL2, ML2, GP1, JL1 |
| 5                  | SM2       | GP2       | PL2       | PM2       | SM1, BK2, PM1, PL1, BK1, MK1, JL2, ML2, GP1, JL1 |
| 4                  | SM2       | GP2       | PL2       | PM2, SM1, BK2, PM1, PL1, BK1, MK1, JL2, ML2, GP1, JL1 |
| 3                  | SM2       | GP2       | PL2, PM2, SM1, BK2, PM1, PL1, BK1, MK1, JL2, ML2, GP1, JL1 |
| 2                  | SM2       | GP2, PL2, PM2, SM1, BK2, PM1, PL1, BK1, MK1, JL2, ML2, GP1, JL1 |
The dendrogram presented in Figure 1 makes it easy to arrange the possible clusters that are formed. To clarify, Table 4 has been compiled with the possible number of clusters that are deemed appropriate from 2 to 6. The two results have led to selecting SM2, GP2, PL2, PM2 stations to be separated into separate clusters. The four always separate themselves if the number of clusters is reduced, with a considerable distance from a group of other stations. Therefore, five is the choice of the number of clusters that are considered sufficient. SM2 station has different characteristics from other clusters in high BOD and COD parameters, while DO parameters tend to be lower. The effect of a lower pH than the others has separated the GP2 station from other clusters while TSS is the parameter that dominantly causes PL2 to be different from the others. BOD and COD also contribute effects, although not as much as TSS. These results are in line with Gonçalves and Alpuim [11], that the variables BOD and COD are strongly related because the cluster structure is very similar. Stations with the worst characteristics in these parameters are assigned to different clusters. Likewise, pH exerts a marked effect and is classified differently because pH classification is more complex and must take into account, for example, the geomorphological characteristics of watersheds. Slightly different at the PM2 station, neither is dominantly different from the others. The causes of the differences are more evenly distributed on almost every parameter.

The results obtained are appropriate with Camara et al. [12] which reported that BOD, COD, DO and pH were parameters that differentiated between clusters of observation objects, in addition to temperature. Similar results were also reported in Alves et al. [13] which presented a moderate association between pH, BOD5, and dissolved oxygen at monitoring water quality of the Sergipe River basin, in both dry and rainy periods. Another study was reported by Liu et al. [14], which clustered river water quality based on different periods. Periods with high river water content give different parameter measurement results. Based on research that the presence of rain on dry agricultural land and residential areas is the largest contributor to the potential for erosion that will cause siltation in the river and reduce the amount of discharge in the river [15].

Considering that the measurement of this parameter was carried out in January, which is the rainy period, it is interesting to investigate whether the clustering results in this study differ in other seasons.

4. Conclusion
Clustering of rivers based on water quality parameters using hierarchical agglomeration technique has been developed. The recommended optimal number of clusters is five, each of which has specific characteristics. The cluster formed shows high homogeneity among cluster members and has high heterogeneity between clusters. Despite the limited amount of data, the findings obtained have explained the clear class differences of rivers at the observed stations. The study needs to be expanded by adding observation stations throughout the province and complementing other water quality parameters to obtain a more comprehensive condition. Likewise, aspects of the development of clustering methods and computational techniques used need to be expanded by utilizing spatial information.

Acknowledgment
This research was supported by the Directorate of Research and Community Service of the Ministry of Education, Culture, Research and Technology with contract No: 187-05/UN7.6.1/PP/2021.

References
[1] de Paul Obade V and Moore R 2018 Environ Int 119 220-231.7.
[2] Ahmed A N, Othman F B, Afan H A, Ibrahim R K, Fai C M, Hossain M S and Elshafie A 2019 J Hydrol 578 124084
[3] Loucks D P and Van Beek E 2017 Springer
[4] Wang Q, Li S, Jia P, Qi C and Ding F 2013 Sci World J 2013 Article ID 231768
[5] Rodriguez M Z, Comin C H, Casanova D, Bruno O M, Amancio D R, Costa L D F and Rodrigues F A 2019 PloS one 14(1) e0210236
[6] S Brinthra Rajakumari and C Nalini 2016 Journal of Chemical and Pharmaceutical Sciences 9(3)
[7] Lee S, Kim J, Hwang J, Lee E, Lee K J, Oh J and Heo T Y 2020 Water 12(9) 2411
[8] Alashwal H, El Halaby M, Crouse J J, Abdalla A and Moustafa A A 2019 Front Comput Neurosci 13 31
[9] Nielsen F 2016 Springer 195–211
[10] Ackerman M and Ben-David S 2016 J Mach Learn Res 17(1) 8182-8198
[11] Gonçalves A M and Alpuim T 2011 Environmetrics 22(8) 933-945
[12] Camara M, Jamil N R B and Abdullah F B 2020 Glob J Environ Sci Manag 6(1) 85-96
[13] Alves J D P H, Fonseca L C, Chielle R D S A and Macedo L C B 2018 RBRH 23
[14] Liu J, Zhang D, Tang Q, Xu H, Huang S, Shang D and Liu R 2021 PloS one 16(1) e0245525
[15] Rezagama A, Sarminingsih A, Zaman B and Handayani D S 2018 IOP Conf. Series: Journal of Physics: Conf. Series 1217 012159