K-MEANS METHOD FOR CLUSTERING WATER QUALITY STATUS ON THE RIVERS OF BANJARMASIN, INDONESIA

Tien Zubaidah, Nieke Karnaningroem and Agus Slamet
Department of Environmental Engineering, Institut Teknologi Sepuluh Nopember, Indonesia
E-Mail: arrasyid.hanif@gmail.com

ABSTRACT

The surface river water quality in Banjarmasin city tends to decline constantly as the result of direct and indirect waste disposal from various human activities along the river body. This study aimed to determine the vulnerability points against pollution in the rivers of Banjarmasin using clustering techniques with K-means algorithm. The parameters observed include Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Suspend Solid (TSS) and Dissolved Oxygen (DO). The data were collected at eight water monitoring stations on various rivers in Banjarmasin city. With the K-means method, four water quality status were clustered. The result showed that 6 stations observed during the period April to October 2016 were catagorized into the heavy polluted cluster with major pollution point of sources came from the domestic and industrial activities.

Keywords: K-Means clustering, water quality, rivers of Banjarmasin.

INTRODUCTION

Banjarmasin and the life of its inhabitants can not be separated from the river. The river in Banjarmasin has various vital functions as a source of drinking water, industrial, agricultural and transportation for its people and plays an important role in supporting the life of aquatic biota. However, the Banjarmasin rivers water quality has been decreased constantly. The decrease of river water quality was strongly influenced by the activities type conducted by residents along the riverbanks that involved the river for their daily activities.

To maintain and achieve river water quality standards that can be sustainable utilized, it is important to make various efforts to control the river water quality. Among the control measures included identifying the rivers critical points in Banjarmasin against pollution. This could be done if only when the water quality standard index has been clustered based on the rivers actual condition. To do that, the modelled water quality measurements needs to be done in order to perform clustering polluted areas prone into certain classes.

The clustering technique was one of the data analysis techniques that allows the retrieval of useful information by grouping or categorizing multidimensional data in clusters. In addition, this technique was essential for data mining systems and to support the decision-making process (Zirnea, Lazar, Foudjo, Vasilache, & Lazar, 2013). The cluster analysis method approach aimed to classified objects in such a way that the distance of each objects to the center of a group in a minimum condition (Zirnea dkk., 2013). K-means clustering technique was one of the grouping techniques that has been widely used in various fields. This technique was based on the distance matrix. K-means algorithm was also one of the simplest, easy to implement and efficient engineering technique of computation and becomes the most popular representative in clustering algorithm (Hartigan & Wong, 1979).

The basic principles of the K-Means algorithm include: assigning k as the number of clusters to be constructed, generating the initial centroid of the cluster randomly, calculating the distance of each data to each cluster center by using Euclidean Distance, grouping each data by distance closest between the data with the center, and the latter determination of the position of the new cluster center (C_{kj}) by calculating the average value of the data that exist in the same cluster center.

The study was expected to classify point of source based on pollutant load, so it will be able to identify the distribution of pollutants along the rivers in Banjarmasin city as well could provide information for local authorities to support the decision making process and policies related to the rivers management.

METHODOLOGY

Site selection

Banjarmasin, as the capital of South Kalimantan province, has been widely known for the existence of its rivers that intersect with each other. The rivers accommodate a variety of heterogeneous activities with various characteristics along the riverbanks (Prayitno, 2012).The research object selection focused on 8 observation points on two main river streams in the city of Banjarmasin, which is Martapura River and Barito River, based on the segmentation result that has been done by The Banjarmasin Environment Agency (see Figure-1). The eight observation points also serve as the location for river quality monitoring station in the city of Banjarmasin.

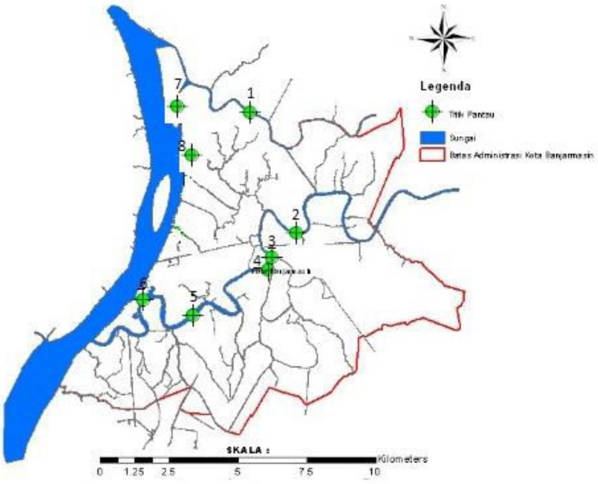

Figure-1. The location and sampling point in the rivers of Banjarmasin.

Framework

The river water quality measurement data that collected by Banjarmasin City Environment Agency during January to December 2016 were used as an analysis material. The study itself involved water quality parameters as a variable, consisted of BOD, COD, TSS and DO for the purpose of classification measurement in this research.The K-Means water quality data clustering was conducted through these following steps:

a. Specifies the number of clusters
b. Determine the centroid randomly before the iteration begins.
c. Perform iterations by calculating the distance of each data object to clusters center with Euclidean distance. Euclidean distance could be measured by the following equation

$$d_{(X,Y)} = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \qquad (1)$$

d. Group the objects by the closest distance between the object and its centroid
e. Determine the new centroid (C_{kj}) by calculating the average values of the objects that existed in the same cluster center.

$$C_{kj} = \frac{x_{1j}+x_{2j}+\cdots+x_{aj}}{a}, j = 1,2,3,\ldots \qquad (2)$$

f. Iteration stoped when the cluster member shares were fixed and there was no change in centroid.

In describing the characteristics of each cluster, the following equations were used:

$$X = \mu + Z.\sigma \qquad (3)$$

In terms of:
X= the samples average (variables in the clusters)
μ = the population average
Z = standardize values
σ = deviation standard

The objects grouping process in this article was conducted with the help of SPSS program.

RESULT AND DISCUSSIONS

By using K-Means clustering algorithm, the minimum, maximum and average values of water quality measurement from 8 observation locations in the rivers of Banjarmasin can be known (see Table-1).

Table-1. Data values description.

Variable values	Min	Max	Avg
TSS April 2016 (X1)	1	29	9,87
BOD April 2016 (X2)	3	8	4,88
COD April 2016 (X3)	8	20	13,25
DO April 2016 (X4)	5	6	5,63
TSS May 2016 (X5)	11	63	27,25
BOD May 2016 (X6)	4	13	8
COD May 2016 (X7)	11	35	21
DO May 2016 (X8)	5	6	5,38
TSS June 2016 (X9)	0	33	16
BOD June 2016 (X10)	2	11	5
COD June 2016 (X11)	6	26	12,75
DO June 2016 (X12)	5	6	5,63
TSS Jule 2016 (X13)	0	39	21
BOD July 2016 (X14)	7	19	14,13
COD July 2016 (X15)	22	49	36,75
DO July 2016 (X16)	4	5	4,75
TSS October 2016 (X17)	7	68	26,5
BOD October 2016 (X18)	3	11	5,38
COD October 2016 (X19)	9	21	14,38
DO October 2016 (X20)	3	6	4,75

The eight water quality monitoring points were grouped into several clusters according to the water quality parameters similarity as measured by the K-Means method. The cluster itself has been divided into the following four categories, the first cluster catagorized as inside the pollution threshold value; the second cluster catagorized as lightly polluted, the third cluster catagorized as moderately polluted; and the forth cluster catagorized as heavily polluted. After the number of clusters was determined, the next step was to determine the centroid randomly before the iteration beguns. Initial centroid can be seen as in Table-2.

Table-2. Initial cluster.

	Cluster			
	1	2	3	4
Zscore: Months of data retrieval	1.396	-1.396	-.698	.698
Zscore: TSS value	.119	-.771	2.7	-1.279
Zscore: BOD value	-.951	-.739	.749	2.026
Zscore: COD value	-.919	-.833	.551	2.109
Zscore: DO value	-.293	1.009	-.293	-1.596
Zscore: Location	-.646	1.508	-1.508	1.077

Then iteration was conducted by calculating the distance of each data object to each cluster center with Euclidean distance. Number of iterations can be seen as in Table-3.

Table-3. Iteration history output.

Iteration	Change in cluster centers			
	1	2	3	4
1	1.157	1.236	1.407	1.256
2	.292	.231	.485	.000
3	.000	.000	.000	.000
a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is,000. The current iteration is 3. The minimum distance between initial centers is 3,871.				

Through the 3 stages of iteration that have been done in the clustering process, it was able to get the right cluster. From Table-3 it can be seen that the minimum distance between the clusters center occurred from the iteration was at 3.871 value. The final result of the clustering process can be seen as in Table-4.

Table-4. Final cluster.

	Cluster			
	1	2	3	4
Zscore: Months of data retrieval	.648	-.829	.174	.581
Zscore: TSS value	.164	-.564	2.217	-.357
Zscore: BOD value	-.237	-.592	.590	1.742
Zscore: COD value	-.233	-.600	.660	1.705
Zscore: DO value	-.386	.684	.032	-.944
Zscore: Location	-.523	.350	-.861	.861

The final cluster centers output mentioned above was still related to the previous data standardization process, which refers to *z-score* with the following conditions:

- A negative value (-) means that the data is below average total
- A positive value (+) means that the data is above the average total

Based on Table-4 and *z-score* provisions, it can be seen that the water quality parameter values in the 1st and 2nd clusters were below average, while the value in 3rd cluster was above the average.The distribution of clustering results with K-Means method which is spread over eight observation points location can be seen in Table-5 below:

Table-5. Number of members in each K-Means cluster.

Cluster	Amount
1	14
2	16
3	4
4	6
Total	40

While the distribution of water quality cluster categories can be seen in Table-6

Table-6. Distribution of each cluster based on monitoring location and data retrieval time.

Locations	Cluster catagories				
	Apr	May	Jun	Jul	Oct
1. PDAM intake	2	3	1	1	3
2. Simpang Arja Port	2	1	1	1	1
3. Belawang port	2	1	1	3	1
4. Anjir muara	2	2	1	4	1
5. Ujung Panti	2	3	2	1	1
6. Kuin Kecil	2	4	2	4	1
7. Muara Kelayan	2	2	2	4	1
8. Basirih	2	2	2	4	4

Remarks:
1: in the threshold cluster
2: lighly polluted cluster
3: moderately polluted cluster
4: highly polluted cluster

From Table-6 it can be seen that in month of July 2016, 4 of 8 observation points (Anjir Muara, Kuin Kecil, Muara Kelayan and Basirih) were clustered in heavy polluted cluster. This condition was suspected as the result of the decrease of river water discharge in Banjarmasin city due to dry season that started in June 2016 (Tarida, 2016).

Those condition were in line with has been stated by Kibena, Nhapi, & Gumindoga (2014) that the pollution increasement along the river during the dry season could be highly attributed to high discharges of effluent from sewage works which cannot be diluted due to low river flows.

In addition, Anjir Muara and Kuin Kecil were located in densely populated areas, with great potential for heavy pollution againts river water. These hypothesis were reinforced by research from (Bu, Liu, Song, & Zhang, 2016)which has confirmed the direct influence of population and population distribution on the quality of river water quality. Studies by (Skovira, 2016) as well (Adebayo & Amer, 2017) also stated that the high dwelling density along the riverbanks was considered responsible for the changes in river water quality, through the faecal coliform increasement as one of the rivers biological indicators.

While at four other observation locations (PDAM intake, Simpang Arja port, Belawang Port and Ujung Panti) were excluded from heavy polluted catagory although at the same time the areas also experienced the dry season. The low distribution of population density in those four locations became the main reason for the exclusion. Meanwhile, Muara Kelayan and Basirih was located in an industrial area that most likely to be moderately to highly polluted. River pollution in industrial areas as described by Lagos (2017) can be caused by the lack of production waste utilization management along the riverbanks.

CONCLUSIONS

The K-means description result againts the water pollution in the rivers of Banjarmasin City showed the clustering of contamination status varies depending on the observation time (seasonal), the activity function characteristics (industrial or domestic) as well as the density of the inhabitants along the riverbanks in Banjarmasin. As the main source of drinking water for its urban population, the segmentation of water quality in the rivers of Banjarmasin city becomes crucial to be grouped according to its pollution level. The clusters will provide an overview of the distribution of river water pollution in the city of Banjarmasin based on selected parameters. It is expected that the clusterification results will serve as a source of consideration for the local governments in producing a more focus and sustainable policies regarding river water resources management.

ACKNOWLEDGEMENT

The research was supported by the Department of Environmental Engineering, Sepuluh Nopember Institute of Technology Surabaya, Indonesia.

REFERENCES

Adebayo S. & Amer R. 2017. Impacts of the Mississippi River spillway opening on faecal coliform concentration in Lake Pontchartrain. River Research and Applications. 33(8): 1327-1335. doi:10.1002/rra.3179.

Bu H., Liu W., Song X. & Zhang Q. 2016. Quantitative impacts of population on river water quality in the Jinshui River basin of the South Qinling Mts., China. Environmental Earth Sciences. 75(4): 292. doi:10.1007/s12665-015-5138-4.

Hartigan J. A.& Wong M. A. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal

Statistical Society. Series C (Applied Statistics). 28(1): 100-108. doi:10.2307/2346830.

Kibena J., Nhapi I.& Gumindoga W. 2014. Assessing the relationship between water quality parameters and changes in land use patterns in the Upper Manyame River, Zimbabwe. Physics and Chemistry of the Earth, Parts A/B/C, 67-69(Supplement C). 153-163. doi:10.1016/j.pce.2013.09.017.

Lagos J. E. S. 2017. Chongqing, a Megalopolis Disconnected with Its Rivers: An Assessment of Urban-Waterside Disconnect in a Chinese Megacity and Proposed Improvement Strategies, Chongqing City as a Case Study. Dalam World Academy of Science, Engineering and Technology, International Journal of Urban and Civil Engineering (Vol. 4). Dubai, UAE: World Academy of Science.

Prayitno B. 2012. A Morphological Analisys on Changing Patern of Banjarmasin Rivercity, Indonesia. Journal of Habitat Engineering and Design. 4(1): 23-32.

Skovira L. 2016. Transforming the Aquatic Urban Landscape: Nutrient Status and Management of Stormwater Basins (Thesis). University of Central Florida, Florida. retreivedfrom http://stars.library.ucf.edu/etd/5593.

Tarida. 2016. BMKG Prediksi Musim Kemarau Mulai Akhir April 2016. access date December 2nd 2017, retrieved from http://dutatv.wixsite.com/duta-tv-banjarmasin/single-post/2016/04/21/Bmkg-Prediksi-Musim-Kemarau-Mulai-Akhir-April-2016.

Zirnea S., Lazar I., Foudjo B. U. S., Vasilache T., &Lazar, G. 2013. Cluster Analysis Based of Geochemical Properties of Phosphogypsum Dump Located Near Bacau City in Romania. APCBEE Procedia, 5(Supplement C), 317-322. doi:10.1016/j.apcbee.2013.05.054.