Water Quality Assessment in the Lam Pa Thao Dam, Chaiyaphum, Thailand with K-Means Clustering Algorithm

2021 Research, Invention, and Innovation Congress: Innovation Electricals and Electronics (RI2C) | 978-1-6654-0300-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/RI2C51727.2021.9559811

Phukkaraphon Ardarsa
Department of Information Technology
Faculty of Informatics, Mahasarakham University
Mahasarakham, Thailand
62011284504@msu.ac.th

Olarik Surinta
Multi-agent Intelligent Simulation Laboratory (MISL)
Department of Information Technology, Faculty of Informatics
Mahasarakham University, Mahasarakham, Thailand
olarik.s@msu.ac.th

***Abstract*— Water resource management is one of the biggest challenges that are being faced, such as a warming climate, arid land, and toxic chemicals in the water. It is essential to deal with water resource management urgently. In this article, researchers mainly focus on monitoring the water quality in the Lam Pa Thao dam, Chaiyaphum, Thailand. The farmer in that area directly affected by the water quality in the dam because they raise fish in floating fish cages. To prevent losses from fish farming, they should have the ability to monitor and control the factors that affect the water quality. As a result, the farmer can monitor the water quality and the monitor system can report to the farmer in time. In this case, to monitor the water quality, researchers designed the buoys, which is the internet of things device, to collect data from the Lam Pa Thao dam. researchers collected the water quality data from January - March 2021, including 13,608 instances. The five important parameters were obtained, including dissolved oxygen, temperature, pH, total dissolved solids, and electric conductivity. Due to the number of parameters, researchers decided not to apply dimension reduction. In these experiments, researchers proposed using K-means clustering algorithms to group the water data into appropriate clusters. For the K-Means algorithm, we calculated the silhouette coefficient to analyze the effectiveness of cluster separation. The best cluster that was grouped using the K-means algorithm achieved the silhouette score of 0.6839. Furthermore, researchers evaluated the K-means algorithm on Charles river and Fitzroy river datasets. It obtained the silhouette score of 0.5489 and 0.6589, respectively.**

Keywords—water quality segmentation, K-means clustering, silhouette coefficient, cluster analysis, internet of things

I. INTRODUCTION

Nowadays, the water supply problem is one of the public issues which affects human life. Water is used in agriculture, consumption, fishery, and public health. Sometimes, water from public health is the source of pathogens and becomes spoilage water.[1] Water is the fundamental factor for human life and the ecosystem, so water management is an urgent issue. [2] It is also the preparation for water problem in the future, For example, solving the water problem in time by using the present technology. [3] Data analysis is one of the technologies helping people to find the hidden data in a specific case and also help us to prove the hypothesis. Water quality analysis is one of the issues which gain attention from several researchers because water still has pollution problems and water supply problems globally. Maintaining water supply to ensure safety for consumption is essential. The fishery and agriculture are also needed high-quality water in their activities. For those reasons, this is an important issue for researchers. [4] To specify the problem and solutions to problems of water quality, water quality data is essential for improving water quality in the long term. [5]

For water data analysis, researchers need to know the water properties in changing environmental conditions and different conditions. Clustering is one of the data analysis techniques. Clustering divides the parameters into groups and parameters are grouped by basic characteristics defined by similarity. [6] The partition of the data is divided in order to define the same elements in the same group for finding the structure of the dataset.[7] This technique is Unsupervised Learning which process objects or values to the appropriate group. Distance of between objects is used as an indicator to identify similarities and differences. Water quality analysis using clustering techniques to find patterns of data distribution is the learning process from samples that is no target value and label value. The datasets are divided into the cluster. The data which contains the same characteristics will be set in the same group.

Water quality segmentation is an important process to make us know the quality of the water source. The quality data also used as a tool to improve water quality, solving the water problem, predicting water quality. This research uses three water quality datasets including, Charles River Buoy Data which was retrieved from the United States Environmental Protection Agency, Fitzroy River Data which is open data recorded by the Queensland government, Australia. The Fitzroy river data, Central coast Queensland was retrieved from the website. The third dataset is Lam Pa Thao dam, Chaiyaphum, Thailand which the researchers have created. The third dataset was created for water quality analysis.

Outline of the paper: This paper is organized in the following way. Section II presents a review of related work. Section III describes the water quality datasets, clustering algorithm, and cluster analysis. The experimental results and conclusion are presented in Section IV and V.

II. RELATED WORK

Carrasco et al. [8] claim that water quality is a sensitive issue. Water quality relates to the characteristics of physical, chemical, and biological elements. Complexities of the area are one thing researchers must concern about. Research in water quality tries to find a model to determine the variables

978-1-6654-0300-9/21/$31.00 ©2021 IEEE

that most affect water quality. The multiple variable analysis help researcher to find relationship and solution for determining the biological, physical and chemical property of water. Water quality evaluation includes several factors such as temperature, pH, transparency, turbidity, nitrates, orthophosphates, phosphorus, total nitrogen, chlorophyll, solar radiation, dissolved oxygen, and microcystins. All of those factors are used as the water quality parameters in Gamboa and Paraiso river in Panama. The research includes three-step. the first step is basic statistical analysis including mean and standard deviation. The second step is HJ-Biplot analysis. The HJ-Biplot technique is the statistical analysis for detecting the correlation of chemical, physical and biological variables. The third step is group analysis. This step identifies the data from the random segmentation of the sampling and identifies affected variables. The segmentation process is calculated by the K-mean method using Biplot data from step two. The result from multiple variable analysis called HJ-Biplot which shows the relationship of chemical, physical, and biological elements. Moreover, the analysis shows two sample data relate to the seasons of the regions. The first group includes variables as follows, pH, transparency, chlorophyll, dissolved oxygen, and temperature. The second group includes variables as follows, nitrates, orthophosphates, turbidity, and others. All parameters normally vary in rainy seasons. The variances cause cyanobacteria and toxigenic in the water.

Hamed [9] claims that water quality monitoring is the priority of surface water protection. There are several methodologies for variable analysis which determine the variability of water quality from different sources. The important part of the variable analysis is the statistical technique and multiple variable statistics. The research decreases the number of variables when testing water in Nile river which are used for Cairo drinking water company. The relationship of each variable is analyzed using Fuzzy C-Means. K-means is used for Segmentation in order to find important factors which change water quality. The result shows that twenty-one water stations can be divided into three segments.

Celestino et al. [10] claims that segmentation by K-means and principal component analysis can be used for water quality analysis and water quality management. However, the previous study claims that K-means which use for finding Euclidean distance is not suitable for high dimensional data. The research uses the algorithm K-means+PCA for the water quality test. The PCA has been used to reduce the dimensions of the data for improving the grouping efficiency of K-means. The huge twenty-eight parameter datasets from hydrogeo chemistry are used. The data are retrieved from 582 water sources of the coastal area of Santo Domingo, Mexico. Segmentation by K-means+PCA using PCA for reducing the dimensions of the data can reduce the attributes of the hydrogeochemical variables from twenty-eight attributes to sixteen attributes. The result shows data are divided into three segments as follows. Segment one includes 160 water sources which are shallow and deep water sources near the coast and near the city. The second segment includes 166 water sources which are mid-level deep and deep water sources located near an industrial area and urban area. The third segment includes 256 water sources which are deep-level water sources near agricultural and urban areas. The result shows the methodology is effective in water source analysis.

Hajigholizadeha and Melesse [11] uses Cluster Analysis (CA) and Discriminant Analysis (DA) in water quality assessment and spatial change assessment in south Florida. The fifteen years dataset from 2000-2014 includes twelve water quality variables from sixteen water stations. The result from sixteen stations shows 35,000 datasets. Water quality from the data can be divided into three segments including low pollution, moderate pollution, and high pollution.

III. Materials and Methods

A. Water Quality Datasets

- *Lam Pa Thao Dam Water Quality Dataset*

The data in this research are collected by buoys equipped with sensors located at Lam Pa Thao Dam, Chaiyaphum, Thailand. Buoys are located at different five points near tilapia cages, the banks of the Lam Pa Thao dam. Total of 13,608 datasets were collected. The data were collected from January – March 2021. The buoy location is shown in Fig. 1.

The water quality of Lam Pa Thao dam, Chaiyaphum, Thailand collected from five floating buoys during January – March 2021 are shown in Fig.2. The data includes five parameters including, dissolved oxygen concentration (DO), temperature, pH, total dissolved solids (TDS), and electric conductivity (EC). The details of the parameters have shown in Table 1.

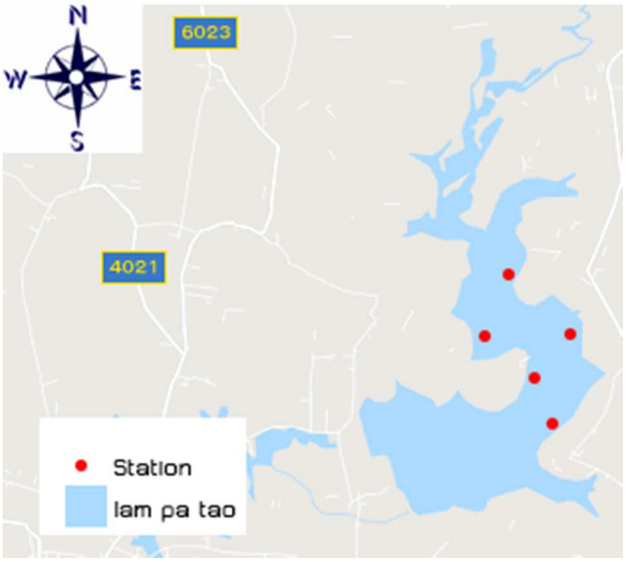

Fig. 1. The position where researchers collect data at Lam Pa Thao dam, Chaiyaphum, Thailand

Fig. 2. The picture of buoys is used as data collector instruments. The buoys are used as data collectors near tilapia fish cages in the dam. The picture also identifies the IoT instruments in each buoy

TABLE I. PARAMETERS WHICH USED TO MEASURE THE WATER QUALITY AND DATA MEASUREMENT UNIT

Parameters	Unit of Measurement
Dissolved Oxygen concentration	milligrams per liter (mg/L)
Temperature	Celsius (°C)
pH	Standard Units
Total Dissolved Solids	Part Per Million (ppm)
Electric Conductivity	percent saturation (%)

- *Charles River Buoy Data*

These data are retrieved from the United States Environmental Protection Agency. They set the buoys at the Lower Charles River for water quality measurement. The water quality data are distributed via the website "https://www.epa.gov/charlesriver/live-water-quality-data-lower-charles-river." Water monitoring buoys are floated at Charles river near the Museum of Science. The sensors are placed one meter below the water surface for collecting data every fifteen minutes. (Table II) shows the 8 collected parameter data including, water temperature, specific conductance, pH, dissolved oxygen concentration, dissolved Oxygen % saturation, turbidity, chlorophyll, and phycoerythrin [12]. This research uses 28,116 Charles river buoy data which are collected during 2018-2019.

TABLE II. PARAMETER FOR WATER MEASUREMENT FROM CHARLES RIVER BUOY

Parameters	Unit of Measurement
Temperature	Celsius (°C)
Specific conductance	millisiemens per centimeter (mS/cm)
pH	Standard Units
Dissolved Oxygen concentration	milligrams per liter (mg/L)
Dissolved Oxygen % saturation	percent saturation (%)
Turbidity	Formazin Nephelometric Units (FNUs)
Chlorophyll	Raw Fluorescent Units (RFUs)
Phycoerythrin	Raw Fluorescent Units (RFUs)

- *Fitzroy River Data*

This data is open data which are collected and distributed by Queensland Government, Australia. The data is water assessment of Fitzroy River, central coast Queensland which collected during 1993-2003. The data include 11,014 datasets. The data are retrieved from https://www.data.qld.gov.au. The data consist of seven parameters including, dissolved oxygen concentration, dissolved oxygen % saturation, pH, salinity, specific Conductance at 25 degrees Celsius, and temperature [13] (Table III)

TABLE III. PARAMETER FOR WATER MEASUREMENT FROM FITZRON RIVER DATA

Parameters	Unit of Measurement
Dissolved Oxygen concentration	milligrams per liter (mg/L)
Dissolved Oxygen (DO)	Percent saturation (%)
pH	Standard Units
Salinity	Practical Salinity Units (PSU)
Specific conductance	millisiemens per centimeter (mS/cm)
Temperature	Celsius (°C)

B. K-Means Clustering Algorithm

K-Means clustering algorithm is an unsupervised learning algorithm developed by MacQueen in 1967. The algorithm uses non-hierarchical cluster analysis which K-Means algorithm automatically divides data into a segment (k) by calculating the distance between Centroid with the target data [14]. It is shown in the following equation.

$$J = \sum_{j=i}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

where k is number of segment (class), n is the number of instances, and c is the (centroid) of the group j.

C. Cluster Analysis

- *The Elbow Method*

The elbow method is the error assessment of the sum of the distance between the centroid and instance for choosing the proper k value. It can be called the optimal cluster number. The appropriate point of the elbow method is where the elbow bend in the graph, as shown in Fig. 3. This point gives the best cluster numbers. [23] [24] Sum of Square Error (SSE) can be calculated as follows.

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (2)$$

where n is the number of instances, y_i is the value of the i^{th} instance, and $\bar{y}$ is the mean of all instances.

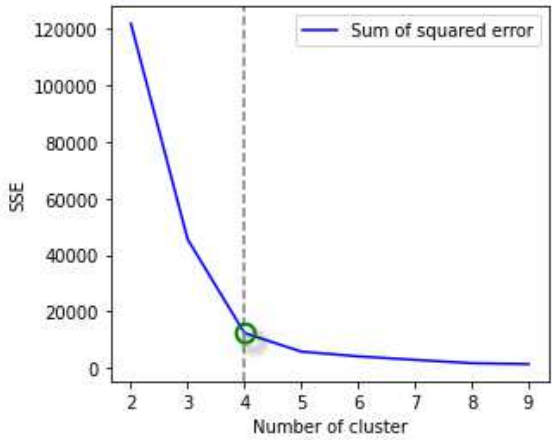

Fig. 3. Illustration of the elbow point

- *Silhouette Coefficient*

Silhouette Coefficient is used for calculating the appropriate number of clusters. Silhouette value range between -1 to +1. The high value indicates the cluster is appropriately divided from the other cluster while a low value means the cluster is divided incorrectly. Silhouette coefficient equations are as follows.

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \text{ if } |C_i| > 1 \quad (3)$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j)$$

where $d(i,j)$ is the distance between instances i and j in cluster (C), k is number of clusters, $a(i)$ is average intra-cluster distance, and $b(i)$ is average inter-cluster distance.

IV. Experimental Results

A. Experimental results on Lum Pa Thao Dam Water Quality Dataset

K-Means algorithm is used in this experiment for dividing the water quality of Lam Pa Thao dam. After that, water data in each cluster is analyzed by the elbow method and Silhouette Coefficient. The result from elbow method shows that the water quality is divided into four clusters. The elbow point is shown in Fig. 4. The calculation result from the Silhouette score shows the appropriate value is 0.6839 which divides the data into 4 clusters as well.

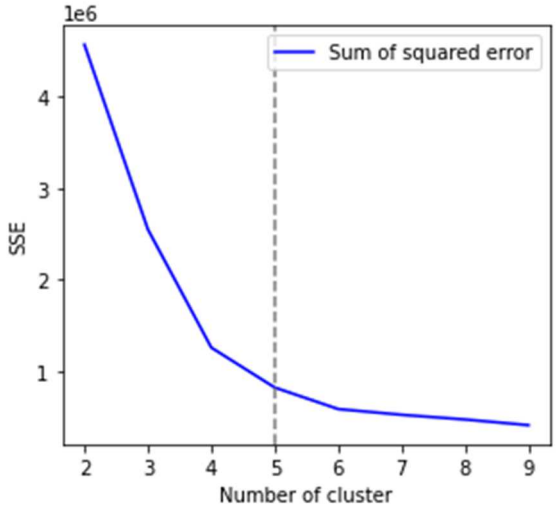

Fig. 4. analysis of suitable segmentation by elbow method

B. Experimental results on Charles River Buoy and Fitzroy River Datasets

The calculation for SSE value with Charles river buoy data and segmentation method by elbow method shows that four clusters are the best point for dividing water quality (See Fig. 5) The result corresponds to the optimal Silhouette value at four clusters as well. The Silhouette value is 0.5489.

Finally, the Fitzroy River Data is segmented by the K-Means algorithm method. The elbow method shows that the optimal number of groups for this data is three clusters, as shown in Fig. 6. The result shows the optimal Silhouette value is 0.6589 which divides the data into three clusters as well.

V. Conclusion

This research aims to study the water quality of Lam Pa Thao dam, Chaiyaphum Province, Thailand because there are a group of tilapia fish farmers around the dam area.

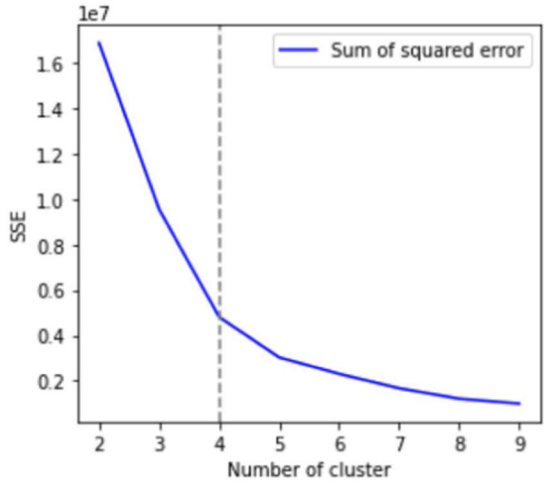

Fig. 5. Optimal number of segmentation analysis by Charles River Buoy Data using elbow method

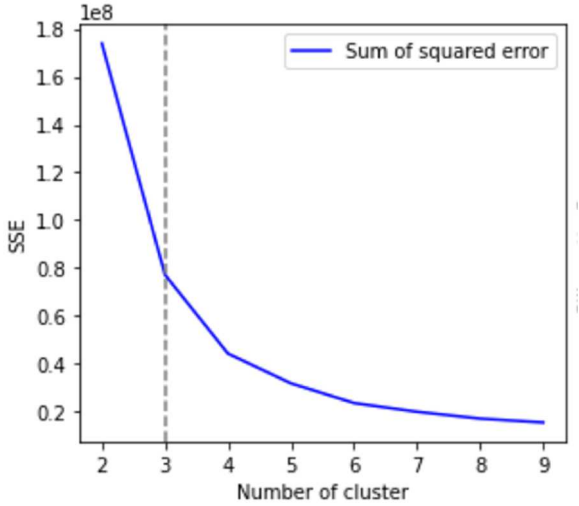

Fig. 6. Analysis of the optimal number of segmentation of Fitzroy River Data using elbow method

In the experiment, the researchers have designed a buoy that can measure the water quality in five parameters including, dissolved oxygen concentration (DO), temperature, pH, total dissolved solids (TDS), and electric conductivity (EC). The buoys are used for collecting data from January to March 2021. 13,608 datasets (instance) are collected from five different locations around the dam area. K-Means algorithm method is used for clustering water quality data. Elbow method and silhouette coefficient are used to analyze the effectiveness of water quality clustering. From the experiment, water quality in Lam Pa Thao dam, Chaiyaphum

is divided into four groups. The optimal point of Silhouette value is 0.6839. The research method is also used with other two datasets including Charles river buoy data and Fitzroy river data. The result shows that Charles river buoy data are divided into four groups, the Silhouette value is 0.5489. the water quality of Fitzroy river is divided into three groups. The Silhouette value is 0.6589.

In Future work, a researcher can test water quality in each cluster to identify the appropriate cluster for tilapia farming.

REFERENCES

[1] K. Rishitha and S. Ullas, "IoT based Automation in Domestic Sewage Treatment Plant to Optimize Water Quality and Power Consumption," in *3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019, pp. 306-310.

[2] L. N. Nthunya, N. P. Khumalo, A. R. Verliefde, B. B. Mamba, and S. D. Mhlanga, "Quantitative analysis of phenols and PAHs in the Nandoni Dam in Limpopo Province, South Africa: A preliminary study for dam water quality management," *Physics and Chemistry of the Earth, Parts A/B/C,* vol. 112, pp. 228-236, 2019/08/01/ 2019.

[3] W.-J. Syu, T.-K. Chang, and S.-Y. Pan, "Establishment of an Automatic Real-Time Monitoring System for Irrigation Water Quality Management," *International Journal of Environmental Research and Public Health,* vol. 17, no. 3, 2020.

[4] M. Tripathi and S. K. Singal, "Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India," *Ecological Indicators,* vol. 96, pp. 430-436, 2019/01/01/ 2019.

[5] P. Luo *et al.*, "Water quality trend assessment in Jakarta: A rapidly growing Asian megacity," *PLOS ONE,* vol. 14, no. 7, p. e0219009, 2019.

[6] F. Ustaoğlu and Y. Tepe, "Water quality and sediment contamination assessment of Pazarsuyu Stream, Turkey using multivariate statistical methods and pollution indicators," *International Soil and Water Conservation Research,* vol. 7, no. 1, pp. 47-56, 2019/03/01/ 2019.

[7] T. Vo-Van, A. Nguyen-Hai, M. V. Tat-Hong, and T. Nguyen-Trang, "A New Clustering Algorithm and Its Application in Assessing the Quality of Underground Water," *Scientific Programming,* vol. 2020, p. 6458576, 2020/03/07 2020.

[8] N. M. Tran, P. Burdejová, M. Ospienko, and W. K. Härdle, "Principal component analysis in an asymmetric norm," *Journal of Multivariate Analysis,* vol. 171, pp. 1-21, 2019/05/01/ 2019.

[9] D. Shen, H. Wu, B. Xia, and D. Gan, "A Principal Component Analysis-Based Dimension Reduction Method for Parametric Power Flow," in *2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, 2020, pp. 1-5.

[10] J. Gan, A. Li, Q. Lei, H. Ren, and Y. Yang, "K-means based on active learning for support vector machine," in *IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, 2017, pp. 727-731.

[11] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *Journal of Physics: Conference Series,* vol. 1361, p. 012015, 2019/11 2019.

[12] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," *IOP Conference Series: Materials Science and Engineering,* vol. 336, p. 012017, 2018/04 2018.

[13] N. U. Godwin Ogbuabor "Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value," *International Journal of Computer Science and Information Technology,* vol. 10, pp. 27-37, 2018.

[14] H. W. Choi, N. M. F. Qureshi, and D. R. Shin, "Comparative Analysis of Electricity Consumption at Home through a Silhouette-score prospective," in *21st International Conference on Advanced Communication Technology (ICACT)*, 2019, pp. 589-591.

[15] U. S. E. P. Agency. (2020, 11/7/2020). Live Water Quality Data for the Lower Charles River. Abailable: https://www.epa.gov/charlesriver/live-water-quality-data-lower-charles-river

[16] Q. G. O. D. Portal. (2014, 12/7/2020). *Fitzroy River (Drainage basin 130)*. Available: https://www.data.qld.gov.au/dataset/ambient-estuarine-water-quality-monitoring-data-1993-to-2013/resource/831dc300-2e2c-4a90-a87e-3dcd74947025