Hydrogeological Cluster Analysis

by Herry Riswandi

Submission date: 22-May-2023 07:07PM (UTC+0700)
Submission ID: 2099179091
File name: tional_Conference_of_Geological_Engineering_Faculty_ICGF_Fix.pdf (1.04M)
Word count: 6485
Character count: 32445

Proceedings of the 5th International Conference of Geological Engineering Faculty, Vol. 1, Issue 2, 2020, pp. 47-61

ISSN 2783-8652 online
DOI: https://doi.org/10.17501/27838652.2020.1204

HYDROGEOLOGICAL CLUSTER ANALYSIS WITH AVERAGE LINKAGE AND WARD METHOD IN THE SOUTHERN SLOPE OF MERAPI MOUNTAIN

Riswandi H[1*], Supandi[2], Sukiyah E[3] and Tania D[4]

[1]*Geological Engineering Department, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia*
[2]*Mining Engineering Department, Institut Teknologi Nasional Yogyakarta, Indonesia*
[3]*Geoscience Department, Universitas Padjadjaran Bandung, Indonesia*
[4]*Geological Engineering Department, Institut Sains & Teknologi Akprind, Indonesia*

Abstract: This study aims to see the cluster analysis process using the average linkage method and the Ward method, and comparing thus results in analysis to clustering several related variables deciding to use the data of depth and hydro chemical character of groundwater. Processing cluster analysis with the average linkage method is pairing objects that combine into one cluster. Then, calculating the two proximity of the object to another variable, the next merging occurs in the most similar clusters than other variables, forming the second cluster. The second combination is to calculate using the average linkage method formula, forming a new distance matrix. The cluster analysis steps with the Ward method starts by the close look at N clusters, which have one respondent for each cluster (all variables consider cluster). The first cluster is formed by selecting two of these N groups, which, when combined, have the smallest value of Error Sum of Squares (SSE). N-1 clusters then consider again to determine which of these two clusters can minimize heterogeneity. Thus, N clusters are systematically reduced by N-1, become N-2, and so on until they become one cluster. The results of clustering the two methods compared with the criteria for standard deviation within groups (S_W) and standard deviation between groups (S_B). The best method has a smaller S_W and S_B ratio. The results showed that the average linkage method and the Ward method have an S_B and S_W ratio value. This result shows that the average linkage method has better performance than the Ward method.

Keywords: gydrogeochemical, groundwater, cluster analysis, linkage method, ward method

Introduction

Cluster analysis clusters similar elements as research objects into different and independent clusters (not related to each other), unlike the discriminant analysis where the cluster determined. Then a discriminant function can be used to determine which element or object should belong to which cluster. While cluster analysis, using specific criteria based on existing data, and indicated by the value of many variables, will form a cluster (Jiang *et al*., 2015). Cluster analysis includes multivariate analysis, but the concept of variate in this technique is different from the concept of variate in other multivariate techniques. In other techniques, variate defines as a linear combination of various variables.

*Corresponding Author's Email: herry.riswandi@upnyk.ac.id

In contrast, in cluster analysis, variation is defined as several variables (which are considered characteristics) comparing an object with another object—an empirical search for variate values not carried out in cluster analysis as in other multivariate techniques. However, cluster analysis's primary purpose is to place a set of objects into two or more clusters based on the similarities of objects based on various characteristics (Anazawa *et al*., 2007). The use of cluster analysis in various geology fields, including morpho-stratigraphy, structural geology, environment, engineering geology, geophysics, and hydrogeology. In the field of geology, cluster analysis can use to assist hydrogeological analysis in determining specific clusters and creating unique programs. In hydrochemistry, cluster analysis can use to identify dissolved elements by type and location. The location study area is on the southern slope of Merapi volcanic mountain, and it is between latitudes 110°22'00" to 110°29'30" N and longitudes 7°35'0" to 7°46'30" E (Figure 1).

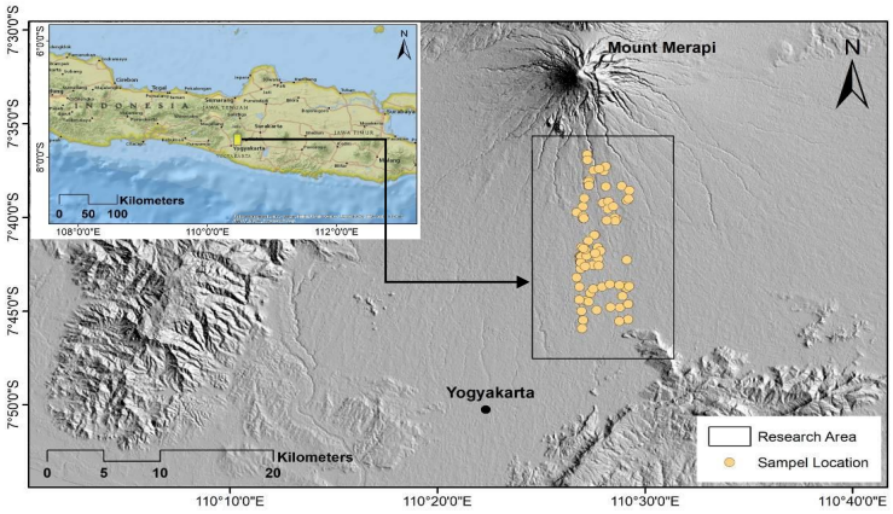

Figure 1: Research area in the southern slope of Mount Merapi in Yogyakarta, Indonesia

Cluster analysis can apply to groundwater by clustering several variables based on deciding to determine the hydrochemical element type. In this research, the hydrochemical elements used are the anion and cation properties of groundwater. Groundwater anions and cations are hydrochemical elements consisting of dissolved groundwater elements and chemical compounds. When someone decides to analyze groundwater's hydrochemistry, they indeed have specific reasons that vary or differ. These reasons are variables or characteristics that will use to compare one respondent with another. There are eight hydrochemical variables used. There is the cation of calium (V_1), natrium (V_2), calsium (V_3), magnesium (V_4), silica (V_5), clorida (V_6), the anion of bicarbonate (V_7), and sulfate (V_8). There are two cluster methods in cluster analysis, the hierarchical method and the non-hierarchical method. Cluster analysis with the hierarchical method is an analysis in which data clustering carried out by measuring each object's proximity, which then forms a dendrogram. There are several types of cluster analysis using the hierarchical method, including the single linkage method, the complete linkage method, the average linkage method, the centroid method, the Ward method, and the median clustering method. In this research, the average linkage method (Demlie *et al*., 2007) and the Ward method (Burghof *et al*., 2017). One reason for using the average linkage method is that it is not a detailed discussion in this research. At the same time, the reason for using the Ward method is

because the Ward method is the best method of cluster analysis with the hierarchical method. After all, this method can minimize the number of squares (error sum of squares, SSE).

This research aims to demonstrate the steps of cluster analysis using the average linkage method. Shows cluster analysis steps using the Ward method and compares the analysis results with the average linkage method for variable data related to using hydrochemical data. After the clustering results obtained, the standard deviation ratio calculate. The ratio obtains from the comparison of the standard deviation in clusters with the standard deviation between clusters. It is useful for knowing which method has the best performance. A right cluster is a cluster that has high similarity between members in one cluster (within the cluster), high heterogeneity between one cluster and another (between clusters). One cluster and other clusters can conclude that a right cluster is a cluster that has members who are as similar as possible to one another but are not very similar to other cluster members. Similarly, in this case, defined as the degree of similarity of characteristics between two data. The smaller the standard deviation ratio within and between clusters, the higher the homogeneity (Gan *et al*., 2018).

Methods

Multivariate analysis is a statistical analysis used to analyze several variables, and these variables are mutually correlated. In general, multivariate analysis divide into two, namely dependency analysis and interdependency analysis. The characteristic of dependency analysis is one or more variables that function as dependent and independent variables, such as multiple linear regression analysis, discriminant analysis, logit analysis, and canonical correlation analysis. The characteristic of interdependency analysis is that all variables are independent. What is included in the interdependence analysis is factor analysis, cluster analysis, and multidimensional scaling. Data in multivariate analysis can express a matrix where there are n objects and p variables. Cluster analysis is a multivariate analysis technique that aims to cluster observational data or variables into clusters. Each cluster is homogeneous following the clustering factors because what is wanted is to get a cluster that is as homogeneous as possible. What is used as the basis for clustering is the similarity of the analyzed scores. Data regarding the size of the similarity can be analyzed with cluster analysis to determine who belongs to which cluster (Mrazovac *et al*., 2013). The cluster analysis steps formulate the problem, choose the distance's size, select the clustering procedure, determine the number of clusters, and interpret the cluster profile (the clusters formed). The essential thing in cluster analysis is selecting variables that will use for clustering (cluster formation), including one or two variables that are irrelevant to the clustering problem that will cause deviations from the clustering results, which are likely to be very useful.

The purpose of cluster analysis is to classify similar objects into the same cluster. Therefore, it requires some measure to know how similar or different objects are. The approach commonly used is to measure the similarities expressed in the distance between pairs of objects. In cluster analysis, there are three measures to measure the similarity between objects: the association's size, the size of the correlation, and the closeness measure. The association measure uses to measure data on a non-metric scale (nominal or ordinal) by taking the correlations coefficient on each object by absolutely negative correlations (Zolekar *et al*., 2020). The correlation measure uses to measure matrix scale data. This measure rarely uses because it focuses on the value of a specific pattern, whereas cluster analysis focuses on the object's size. The similarity between objects can see from the correlation coefficient

between pairs of objects measured using several variables. The Euclidean distance measures the sum of the squares of the difference in values for each variable (1).

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (1)$$

d_{ij} is the distance between the ith object and the nth object, p is the number of cluster variables, x_{ik} is the data from the ith subject in the k-variable, and x_{jk} is the data from the j subject in the k-variable. Squared Euclidean distance is a variation of the distance Euclidean. If its roots at the Euclidean distance, then the root is removed at the Squared Euclidean distance (2).

$$d_{ij} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2 \quad (2)$$

Cluster analysis The process of forming clusters can do in two ways: hierarchical and non-hierarchical methods. The hierarchical method consists of the agglomerative method and the divisive method. The agglomerative method consists of 3 methods; there are the linkage method, the variance method, and the centroid method. The linkage consists of the single linkage method, complete linkage, and average linkage. Meanwhile, the variance method consists of the Ward method. The non-hierarchical method consists of three methods: the sequential threshold method, the parallel method, and the optimizing partitioning method (Figure 2).

The hierarchical method is a cluster analysis method that forms a certain level, such as in a tree structure because the clustering process carries out in stages. The results of clustering using the hierarchical method can present in the form of a dendrogram. The dendrogram is a visual representation of the steps in a cluster analysis that shows how the clusters formed and each step's distance coefficient value. The figure on the right is the object of research, where these objects are connected by lines with other objects so that they will form a cluster in the end. The methods that can use in the hierarchical method are the agglomerative method and the divisive method. The agglomerative method starts with assuming that each object is a cluster. Then the two objects with the closest distance are combined into one cluster.

Furthermore, the third object will join the existing cluster or with other objects and form a new cluster while still considering the proximity between objects. The process will continue until, finally, a cluster consisting of all objects created. Several agglomerative methods, like the single linkage method, the complete linkage method (farthest-neighbor method), centroid method, average linkage method, and the Ward method (Figure 2). The process in the divisive method is opposite to the agglomerative method. This method starts with one large cluster that includes all objects of observation. Then, gradually objects that have a large enough dissimilarity will be separated into different clusters. The process carries out so that the desired number of clusters is formed, such as two clusters. The Single Linkage method used to determine the distance between clusters using the single linkage method can be done by looking at the distance between the two existing clusters and then selecting the closest distance or near-neighbor rule (Johnson & Wichern, 1992).

In this case, the quantities $d_{(ik)}$ and $d_{(jk)}$, respectively, are the shortest distance between clusters I and K and clusters J and K. The clustering results using the single linkage method can be displayed graphically in a dendrogram or tree diagram. The branches on the tree represent the number of clusters. In the complete linkage method, the distance between clusters determined by the farthest-neighbor distance between two objects in different clusters where $d_{(ik)}$ and $d_{(jk)}$ each is the distance between the most distant members of clusters I and J and clusters J and K (Johnson & Wichern, 1992).

The centroid method is the average of all objects in the cluster. In this method, the distance between the clusters is the distance between centroids. The new centroid calculates whenever an object is combined so that each time the members increase, the centroid will change. In the centroid method, the distance between the clusters is the distance between the centroids. The centroid is the average of all the members in the cluster. When objects are combined, new centroids calculates so that every time there is an addition of members, the centroid will change as well (Johnson & Wichern, 1992). In the average linkage method, the distance between two clusters is considered the average distance between all members in one cluster and all other clusters (3).

$$d_{(ij)k} = \frac{\Sigma a \Sigma b \, d_{ab}}{N_{ij} N_{ik}} \quad (3)$$

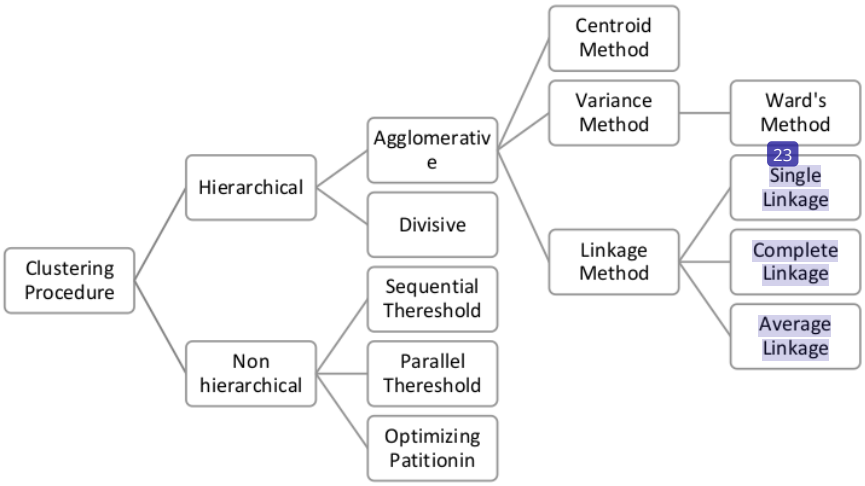

Figure 2: The classification of cluster analysis procedures.

The variant method aims to obtain clusters that have the smallest possible cluster internal variance. The commonly used variance method is the Ward method, where the average for each cluster calculates. Then, calculate the Euclidean distance between each object and the average value, then calculated all the distances. The two clusters with the smallest increase in the 'sum of squares in the cluster' combined at each stage. Ward method is a cluster formation method based on the loss of information due to the merging of objects into clusters. It measured using the sum of the squared deviations in the cluster means for each observation—the sum of squares (SSE) error used as an objective function. Two objects will combine to have the smallest objective function among the existing possibilities (4).

$$SSE = \sum_{j=1}^{p} \left(\sum_{i=1}^{n} x_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_{ij} \right)^2 \right) \quad (4)$$

The main problem in cluster analysis is determining how many clusters there are. There are no standard rules for determining how many clusters there are. However, some pointers can use. For example, suppose the purpose of clustering is to identify element segments. In that case, management may want a certain number of clusters (say 3, 4, or 5 clusters), and the relative size of the clusters should be useful. The interpretation stage includes testing each cluster that forms to provide a name or

description precisely as a description of the cluster's nature, explaining how they can be relevant in each dimension. When starting the interpretation process, each cluster's average (centroid) uses for each variable.

This research use data collected form of field data and laboratory data. Field data consists of surface data and subsurface data. Each of them contains sample data directly observed in the field and samples taken to the laboratory to analyze groundwater's physical and chemical properties—the hydrogeological and statistical data in the database created in different layers of information. After the hydrogeological data compiled, the validity test carries because the data is valid if the data reveals something measured by the data. The validity test uses to measure whether the measurements and observations made are relevant or not. The validity test performed using the Pearson formula in SPSS.

In this study, sampling used the multivariate analysis approach (Tabachnick & Fidell, 2013). Sampling using the Tabachnick & Fidell technique is the number of independent variables multiplied by a weight of 10-25. The number of independent variables in this study was 8, so the number of samples needed was 80-200. In this study, the weight chosen was 25, so that the number of samples to be taken was 89 samples for groundwater depth.

After the data is collected, it is processed to provide an overview of the existing problems. The data processing stages are; (i) editing; the initial stage of data analysis is to edit the data collected. The data editing process that aims to make the data later will be analyzed is accurate and complete. (ii) After the editing stage, encoding is complete; the next step needs to do is coding. Coding is the provision of codes; each data includes assigning categories for the same type of data. The code is a symbol in the form of letters or numbers to provide identity data. (iii) Tabulation is the process of placing coded data in table form according to the needs of the analysis. (iv) Data Processing, after all the data goes through the editing and coding process, the next step is to process the data to be analyzed. In this study, data processing was carried out with software, namely SPSS, using the average linkage method and the Ward method.

The data processing stages from the research carried out are as follows; (i) check the answers' completeness. At this stage, obtained data will check again to determine the answers to the required data obtained entirely. (ii) Compilation of the data obtained so that it is easy to analyze or use at a later stage. (iii) Classification of data using cluster analysis, starting from selecting the distance's size, choosing the clustering procedure, determining the number of clusters, and interpreting the cluster profile formed.

Result

The data in this discussion is to classify hydrochemical related to the reasons for selecting groundwater quality. The data that will use is 89 groundwater data—obtained data by analyzing groundwater anions and cations in the laboratory. After the data is collected, the data entered into the analysis results table with 89 data. In this table, the number of respondents is 1, 2, 3, and continue. There is the cation of calium (V_1), natrium (V_2), calsium (V_3), magnesium (V_4), silica (V_5), clorida (V_6), the anion of bicarbonate (V_7), and sulfate (V_8).

Variable standardization carries out when there are significant unit differences between the variables studied. However, if the data collected does not have unit variability, then the cluster analysis process can be carried out immediately without standardization. Because the data in *Table 1* has the same unit scale, data standardization not use in this study. Distance size using the Euclidean distance means the

distance between object *i* and *j*, a pair of objects to be measured for similarity. The Euclidean distance formula follows equation (2).

Eighty-nine data will measure for similarity. Given an example of a calculation to calculate the distance between respondents 1 and 2, respondents 1 and 3, respondents 2 and 3. The three respondents compared using eight variables to find the two most similar hydrochemical data among the three. Table 2 calculates respondents 1 and 2 results in an Euclidean distance of 10.283 or produces a Squared Euclidean distance of 3.207. In *Table 3*, the calculation respondents 1 and 3 result in a Euclidean distance of 0.080 or a Squared Euclidean distance of 0.282. In *Table 4*, the calculation respondents 2 and 3 results in a Euclidean distance of 10.508 or a Squared Euclidean distance of 3.242.

It can see from the calculation that the closest pair of the three respondents, according to Euclidean distance, are respondents 1 and 3 because their scores are the lowest, namely 0.080, or according to the Squared Euclidean distance, namely 0.282. The lower the distance score, the closer the paired respondents.

Table 1: Sample value for 89 respondents of hydrochemical data.

Value Sample Respondent	Groundwater Depth (meter)	Element ($^{meq}/_{L}$)							
		V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
		K^+	Na^+	Ca^{2+}	Mg^{2+}	SiO_2	Cl^-	HCO_3^-	SO_4^{2-}
Minimum	0.200	0.051	0.174	0.159	0.120	0.116	0.017	0.649	0.249
Maximum	76.000	0.921	7.090	1.532	1.849	1.708	1.453	8.362	9.910
Median ($\bar{x}$)	9.933	0.319	1.714	0.627	0.785	0.917	0.308	3.493	1.483

Table 2 : Calculation of the proximity of respondents 1 and 2.

Respondent	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	Total
	K^+	Na^+	Ca^{2+}	Mg^{2+}	SiO_2	Cl^-	HCO_3^-	SO_4^{2-}	
1	0.358	3.958	0.619	0.637	1.106	0.282	3.671	0.708	
2	0.409	7.090	0.576	0.398	0.645	0.465	4.078	0.750	
$(V_{ik} - V_{jk})$	0.051	3.132	-0.043	-0.239	-0.461	0.183	0.406	0.042	
$(V_{ik} - V_{jk})^2$	0.003	9.808	0.002	0.057	0.213	0.034	0.165	0.002	10.283
Using equation (1)									3.207

Table 3: Calculation of the proximity of respondents 1 and 3.

Respondent	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	Total
	K^+	Na^+	Ca^{2+}	Mg^{2+}	SiO_2	Cl^-	HCO_3^-	SO_4^{2-}	
1	0.358	3.958	0.619	0.637	1.106	0.282	3.671	0.708	
3	0.256	3.915	0.536	0.756	1.109	0.339	3.874	0.750	
$(V_{ik} - V_{jk})$	-0.102	-0.043	-0.083	0.119	0.002	0.056	0.203	0.042	
$(V_{ik} - V_{jk})^2$	0.010	0.002	0.007	0.014	0.000	0.003	0.041	0.002	0.080
Using equation (1)									0.282

Table 4: Calculation of the proximity of respondents 2 and 3.

Respondent	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	Total
	K^+	Na^+	Ca^{2+}	Mg^{2+}	SiO_2	Cl^-	HCO_3^-	SO_4^{2-}	
2	0.409	7.090	0.576	0.398	0.645	0.465	4.078	0.750	
3	0.256	3.915	0.536	0.756	1.109	0.339	3.874	0.750	
$(V_{ik} - V_{jk})$	-0.153	-3.175	-0.040	0.358	0.464	-0.127	-0.203	0.000	
$(V_{ik} - V_{jk})^2$	0.024	10.083	0.002	0.128	0.215	0.016	0.041	0.000	10.508
Using equation (1)									3.242

This research uses two clustering methods: the average linkage method and the Ward method.

1. Clustering using the average linkage method

The clustering process using the average linkage method is through the following steps; (i) the pairs of close together respondents are combined into one cluster, namely respondents 79 and 80 (*Table 5*). (ii) Calculate the distance of respondents 79 and 80 who joined into one cluster with other respondents. (iii) The next merging occurs in the most similar clusters, thus forming the second cluster. Then calculated using the formula (3) so that a new distance matrix formed. (iv) Repeating steps i and iii, N-1 times, where N is the number of objects or respondents.

Table 5: Agglomeration schedule with Average Linkage Method.

Stage	Cluster Combined		Coefficients	Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2			Cluster 1	Cluster 2	
1	79	80	0.000	8	73	78	0.047
2	74	87	0.027	9	9	13	0.054
3	45	47	0.030	10	66	69	0.060
4	66	67	0.033	11	70	73	0.065
5	4	8	0.038	12	55	56	0.069
6	70	83	0.041	...	...	...	...
7	11	12	0.046	88	1	23	64.144

The clustering process can also do using SPSS, namely at the agglomeration stage, as in *Table 5*. Agglomeration Schedule using the average linkage method in *Table 5* results from clustering using the average linkage method. After the Squared Euclidean distance measured the distance between variables, clustering carries out in stages. In the first step, a cluster formed consisting of respondents number 70 and 83 with a distance of 0.041 (given in the coefficients column). Because the agglomeration process starts from the two closest objects, this distance is the closest of the many combinations of the 89 objects. Then, in the next stage column, the number 11 is a column that shows the stages were other respondents combined with the newly formed cluster. The newly formed cluster means that the next clustering process carries out at stage 11. In the second step, at stage 11, it can see that 70 respondents form clusters with 73 respondents who have a distance of 0.065. This distance is the minimum distance of the last object that joins the two previous objects. The clustering process above can also illustrate in the form of a dendrogram (Figure 3). The dendrogram read from left to

right, where the vertical lines show the clusters joined together, while the lines on the scale show the distance of the clusters joined together.

2. Clustering using the average Ward method

The Ward method's clustering process is through stages following; (i) starting with paying attention to N clusters with one respondent per cluster (all respondents are assigned a cluster). SSE will be zero for the first stage because each respondent will form a cluster. (ii) The first cluster form by selecting two of the N clusters with the smallest SSE value. The smallest SSE value is in line with its objective function, namely, minimizing heterogeneity. The SSE using formula (4). (iii) N-1 cluster clusters consider again to determine which two of these clusters can minimize heterogeneity. Thus, N clusters were systematically reducing by N-1. (iv) Repeating steps 2 and 3 until one cluster obtained or all respondents combine into one cluster.

Like clustering with the average linkage method, clustering with the Ward method can also do with SPSS, which is at the agglomeration stage shown in *Table 6*. The Agglomeration Schedule in *Table 6* is the result of clustering using the Ward method. After the euclidean distance measured the distance between variables, clustering carries out in stages. The clustering process above can also illustrate in the form of a dendrogram (Figure 3). The dendrogram read from left to right, where the vertical lines show the clusters joined together, while the lines on the scale show the distance of the clusters joined together.

Table 6: Agglomeration schedule with Ward Method.

Stage	Cluster Combined		Coefficients	Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2			Cluster 1	Cluster 2	
1	79	80	0.000	8	73	78	0.131
2	74	87	0.014	9	9	13	0.158
3	45	47	0.029	10	66	69	0.193
4	66	67	0.045	11	55	56	0.228
5	4	8	0.064	12	20	21	0.266
6	70	83	0.085	...	...	...	...
7	11	12	0.108	88	1	14	515.992

3. Determine the number of clusters and their members

The agglomeration process is complicated, especially in the calculation of coefficients involving multiple respondents and increasing. The agglomeration process will ultimately unite all respondents into one cluster. In the process, several clusters are generated with each member, depending on the number of clusters formed. The cluster analysis only shows the cluster members for a certain number of clusters, not how many clusters formed. This research uses cluster membership with 4 clusters because it expects that the results obtained are more accurate and closer to the actual situation.

Details of the number of clusters and members formed show in the SPSS cluster membership output table using the average linkage and Ward method. From the SPSS cluster membership can be concluded that the members of each cluster in *Table 7*:

- Output table with the average linkage method, it is known that cluster 1 consists of 82 respondents, cluster 2 consists of 4 respondents, cluster 3 consists of 1 respondent, and cluster 4 consists of 2 respondents.
- Output table with the Ward method, it is known that cluster 1 consists of 13 respondents, cluster 2 consists of 62 respondents, cluster 3 consists of 12 respondents, and cluster 4 consists of 2 respondents.

Table 7: Cluster members using the average linkage method and with the Ward method

No. Cluster	Cluster members	
	using the average linkage method	using the Ward method
Cluster 1	1, 3-5, 7-9, 13-22, 24-25, 27-89	1-13
Cluster 2	2, 6, 11-12	14-16, 18, 20-22, 24-25, 27-28, 31-34, 42, 44-56, 58-89
Cluster 3	10	17, 19, 29-30, 35-41, 43, 57
Cluster 4	23, 26	23, 26

Figure 3: Dendrogram with Average Linkage and Dendrogram with Ward Method.

The number of respondents who enter the 4 clusters based on their element can see in the bar chart *Figure 4*.

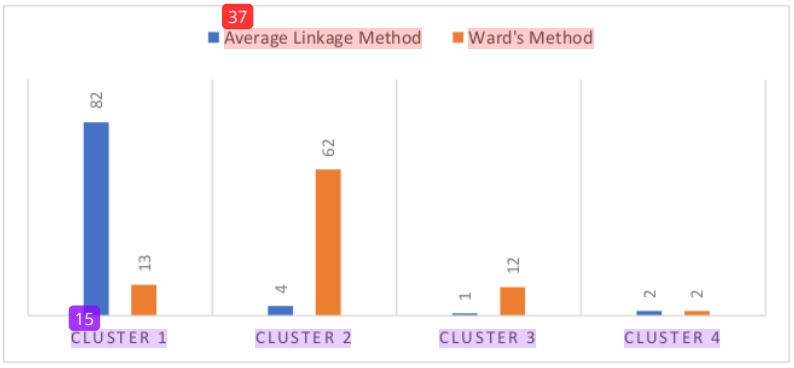

Figure 4: Bar chart of the number of elements on the average linkage and Ward method

4. Cluster Interpretation with average linkage method

After determining the number of clusters and their members, the next step is cluster interpretation. It can interpret the cluster profile by using each cluster's average for each variable (centroid). The cluster profile using the average linkage method (*Table 8*) and the centroid values for each variable are in clusters. From *Table 8*, cluster 1 and cluster 3 can see that V_7 has a high centroid value than other variables, so V_7 (HCO_3^-) is the highest reason respondents in cluster 1 and cluster 3 bicarbonate are a dominant element. Cluster 2 shows that V_2 has a high centroid value than other variables, so V_2 (Na^+) is the highest reason respondents in cluster 2 natrium are dominant elements. Cluster 4 shows that V_8 has a high centroid value than other variables, so V_8 (SO_4^{2-}) is the highest reason respondents in cluster 4 sulfate are dominant elements.

Table 8: Value of centroid cluster 1 to 4 with the Average Linkage Method.

Centroid Value	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
	K^+	Na^+	Ca^{2+}	Mg^{2+}	SiO_2	Cl^-	HCO_3^-	SO_4^{2-}
Cluster 1	0,319	1,451	0,611	0,789	0,922	0,306	3,436	1,374
Cluster 2	0.345	6.829	0.571	0.567	0.810	0.413	4.436	0.479
Cluster 3	0.256	3.828	0.477	0.756	0.962	0.183	8.362	0.250
Cluster 4	0.320	1.196	1.495	1.055	0.891	0.233	1.500	8.557

5. Cluster Interpretation with Ward's method

After determining the number of clusters and their members, the next step is cluster interpretation. It can interpret the cluster profile by using each cluster's average for each variable (centroid). The cluster profile using the Ward method (*Table 9*) and the centroid values for each variable are in clusters.

Table 9, cluster 1 can see that V_2 has a high centroid value compared to other variables, so V_7 (Na^+) is the highest reason respondents in cluster 1 natrium are the dominant element. Cluster 2 shows that V_7 has a high centroid value than other variables, so V_7 (HCO_3^-) is the highest reason respondents in cluster 2 bicarbonate are dominant elements. Cluster 3 and cluster 4 show that V_8 has a high centroid value than other variables, so V_8 (SO_4^{2-}) is the highest reason respondents in cluster 3 and cluster 4 sulfate are dominant elements.

Table 9: Value of centroid cluster 1 to 4 with the Ward Method.

Centroid Value	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
	K^+	Na^+	Ca^{2+}	Mg^{2+}	SiO_2	Cl^-	HCO_3^-	SO_4^{2-}
Cluster 1	0.293	4.781	0.529	0.664	0.958	0.298	4.472	0.458
Cluster 2	0.321	1.207	0.608	0.790	0.907	0.307	3.664	1.152
Cluster 3	0.338	1.101	0.684	0.839	0.929	0.334	2.021	2.972
Cluster 4	0.320	1.196	1.495	1.055	0.891	0.233	1.500	8.557

6. Determining the Merits of the Clustering Method with Standard Deviation

The two clustering methods performance determine the standard deviation criteria used: the average standard deviation in the cluster (S_W) and the standard deviation between clusters (S_B). The best method has the smallest standard deviation ratio (S_W) and inter-cluster standard deviation (S_B). The smaller the Sw value and the greater the SB value, the better the method is, meaning high homogeneity (Bunkers *et al*., 1996).

Determine standard deviation (S_W) in clusters and between clusters in the average linkage method (*Table 10*). The standard deviation where the median value of cluster 1 x_i is 1.151 and S_W is 0.258, cluster 2 x_i is 1.806 and S_W is 0.019, cluster 3 x_i is 0, because there is only 1 object in the cluster, then S_W has no effect (value 0), and cluster 4 x_i is 1.906 and S_W is 0.225.

Determine standard deviation (S_W) in clusters and between clusters in the Ward method (*Table 10*). The standard deviation where the median value of cluster 1 x_i is 1.557 and S_W is 0.223, cluster 2 x_i is 1.119 and S_W is 0.279, cluster 3 x_i is 1.152 and S_W is 0.118, and cluster 4 x_i is 1.906 and S_W is 0.225.

Table 10 : Median variable for each respondent by average linkage and Ward method.

No. Cluster	Respondent median variable					
	using the average linkage method			using the Ward method		
	total	median (x_i)	S_W	total	median (x_i)	S_W
Cluster 1	94.385	1.151	0.258	20.239	1.557	0.223
Cluster 2	7.225	1.806	0.019	68.275	1.119	0.279
Cluster 3	0.250	0.250	0	14.980	1.152	0.118
Cluster 4	3.811	1.906	0.225	3.811	1.906	0.225
		$\bar{x}$i = 1.278	$\bar{x}$ S_W = 0.126		$\bar{x}$i = 1.434	$\bar{x}$ S_W = 0.211

- Standard deviation between clusters (S_B) using the average linkage method using the equation:

$$S_B = [(4-1)^{-1} \sum_{k=1}^{4} (\bar{x}_k - \bar{x})^2]^{\frac{1}{2}} \quad (3)$$

$$S_B = \left(\frac{(1.151 - 1.278)^2 + (1.806 - 1.278)^2 + (0.250 - 1.278)^2 + (1.906 - 1.278)^2}{4-1} \right)^{\frac{1}{2}}$$

$$S_B = \left(\frac{1.746}{3}\right)^{\frac{1}{2}} = (0.582)^{\frac{1}{2}} = 0.763$$

$$Rasio = \frac{S_W}{S_B} = \frac{0.126}{0.763} = 0.165$$

So the standard deviation value between clusters using the average linkage method is 0.763, and the ratio between standard deviation within and between clusters is 0.165.

- The standard deviation between clusters (S_B) using the Ward method is using equation (9).

$$S_B = \left(\frac{(1.557-1.434)^2 + (1.119-1.434)^2 + (1.152-1.434)^2 + (1.906-1.434)^2}{4-1}\right)^{\frac{1}{2}}$$

$$S_B = \left(\frac{0.422}{3}\right)^{\frac{1}{2}} = (0.141)^{\frac{1}{2}} = 0.375$$

$$Rasio = \frac{S_W}{S_B} = \frac{0.211}{0.375} = 0.562$$

So the standard deviation value between clusters using the Ward method is 0.375, and the ratio between standard deviation within and between clusters is 0.562.

The value of the standard deviation ratio in clusters and the standard deviation between clusters shows that the average linkage method has better performance than the Ward method. The difference because the average linkage method has smaller ratio values, which is 0.165, compared to the Ward method's ratio value of 0.562.

Conclusions

Regarding the application of cluster analysis with the average linkage method and the Ward method for groundwater hydrochemical respondent data, it can conclude that there are differences in the cluster analysis steps with the average linkage method and the Ward method.

The value of the standard deviation ratio in clusters and between clusters can assess the comparison of the clustering method's performance. The best method is the method that has the smallest standard deviation ratio within and between clusters. The ratio value for groundwater hydrochemical respondent data shows that the average linkage method performs better than the Ward method.

References

Anazawa, K., Sakamoto, H., & Tomiyasu, T. (2007). Influence of Ignimbrite on the Chemistry of River Water in Shirasu Plateau, Japan. *Hydrogeology Journal,* 15(1), 409–417. DOI:10.1007/s10040-006-0070-z

Bunkers, M. J., James R. Miller, J., & DeGaetano, A. T. (1996). Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique. *Journal of Climate,* 9(1), 130-146. doi:doi.org/10.1175/1520-0442(1996)009<0130:DOCRIT>2.0.CO;2

Burghof, S., Gabiri, G., Stumpp, C., Chesnaux, R., & Reichert, B. (2017). Development of a Hydrogeological Conceptual Wetland Model in the Data-Scarce North-Eastern Region of Kilombero Valley, Tanzania. *Hydrogeology Journal,* 26(1), 267-284. DOI:10.1007/s10040-017-1649-2

Demlie, M., Wohnlich, S., Wisotzky, F., & Gizaw, B. (2007). Groundwater Recharge, Flow, and Hydrogeochemical Evolution in a Complex Volcanic Aquifer System, Central Ethiopia. *Hydrogeology Journal,* 15(1), 1169–1181. DOI:10.1007/s10040-007-0163-3

Gan, Y., Zhao, K., Deng, Y., Liang, X., Ma, T., & Wang, Y. (2018). Groundwater Flow and Hydrogeochemical Evolution in the Jianghan Plain, Central China. *Hydrogeology Journal,* 26(1), 1609–1623. DOI:10.1007/s10040-018-1778-2

Jiang, Y., Guo, H., Jia, Y., Cao, Y., & Hu, C. (2015, June). Principal Component Analysis and Hierarchical Cluster Analyses of Arsenic Groundwater Geochemistry in the Hetao basin, Inner Mongolia. *Geochemistry,* 75(2), 197-205. DOI:10.1016/j.chemer.2014.12.002

Johnson, R. &. (1992). Applied Multivariate Statistical. New Jersey: Prentice-Hall International.

Mrazovac, S., Mirjana, V.-M., Matić, I., & Marić, N. (2013, June). Multivariate Statistical Analyzing of Chemical Parameters of Groundwater in Vojvodina. *Geochemistry,* 73(2), 217-225. DOI:10.1016/j.chemer.2012.11.002

Tabachnick, B. G., & Fidell, L. S. (2013). Using Multivariate Statistics. California: Pearson.

Zolekar, R. B., Todmal, R. S., Vijay S. Bhagat, S. A., Korade, M. S., & Das, S. (2020, June 09). Hydrochemical Characterization and Geospatial Analysis of Groundwater for Drinking and Agricultural Usage in Nashik district in Maharashtra, India. Environment, Development, and Sustainability, 1-20. DOI:10.1007/s10668-020-00782-2

Hydrogeological Cluster Analysis

ORIGINALITY REPORT

8%	3%	4%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Help University College Student Paper	<1%
2	Submitted to HELP UNIVERSITY Student Paper	<1%
3	www.scholink.org Internet Source	<1%
4	Juan Manuel Vilar. "Classifying Time Series Data: A Nonparametric Approach", Journal of Classification, 04/22/2009 Publication	<1%
5	Submitted to Syiah Kuala University Student Paper	<1%
6	Submitted to Coventry University Student Paper	<1%
7	repository.ub.ac.id Internet Source	<1%
8	Submitted to Nottingham Trent University Student Paper	<1%

9	Submitted to Universitas Muhammadiyah Surakarta Student Paper	<1 %
10	worldwidescience.org Internet Source	<1 %
11	Andrea Alonso, Andrés Monzón, Iago Aguiar, Alba Ramírez-Saiz. "Explanatory Factors of Daily Mobility Patterns in Suburban Areas: Applications and Taxonomy of Two Metropolitan Corridors in Madrid Region", ISPRS International Journal of Geo-Information, 2023 Publication	<1 %
12	Submitted to Far Eastern University Student Paper	<1 %
13	Ufuk Bolukbas, Ali Fuat Guneri. "Knowledge-based decision making for the technology competency analysis of manufacturing enterprises", Applied Soft Computing, 2017 Publication	<1 %
14	Submitted to University of Hong Kong Student Paper	<1 %
15	Submitted to University of Wales, Bangor Student Paper	<1 %
16	proceeding.researchsynergypress.com Internet Source	<1 %

17	Submitted to Colorado Technical University Online Student Paper	<1 %
18	Submitted to UI, Springfield Student Paper	<1 %
19	digilib.uin-suka.ac.id Internet Source	<1 %
20	Chu Wu, Chen Fang, Xiong Wu, Ge Zhu, Yuzhe Zhang. "Hydrogeochemical characterization and quality assessment of groundwater using self-organizing maps in the Hangjinqi gasfield area, Ordos Basin, NW China", Geoscience Frontiers, 2021 Publication	<1 %
21	Submitted to Graz University of Technology Student Paper	<1 %
22	Submitted to University of Bath Student Paper	<1 %
23	www.idosi.org Internet Source	<1 %
24	Jing Yang, Ming Ye, Zhonghua Tang, Tian Jiao, Xiaoyu Song, Yongzhen Pei, Honghua Liu. "Using cluster analysis for understanding spatial and temporal patterns and controlling factors of groundwater geochemistry in a regional aquifer", Journal of Hydrology, 2020	<1 %

Publication

25	Katsuro Anazawa, Hayao Sakamoto, Takashi Tomiyasu. "Influence of ignimbrite on the chemistry of river water in Shirasu plateau, Japan", Hydrogeology Journal, 2006 Publication	<1 %
26	Kuo, R.J.. "Integration of self-organizing feature map and K-means algorithm for market segmentation", Computers and Operations Research, 200209 Publication	<1 %
27	M. Salmenkivi. "Multivariate Analysis of Finnish Dialect Data An Overview of Lexical Variation", Literary and Linguistic Computing, 05/02/2007 Publication	<1 %
28	Michalis Vazirgiannis, Maria Halkidi, Dimitrios Gunopulos. "Uncertainty Handling and Quality Assessment in Data Mining", Springer Science and Business Media LLC, 2003 Publication	<1 %
29	Paul Perco. "Transforming omics data into context: Bioinformatics on genomics and proteomics raw data", Electrophoresis, 07/2006 Publication	<1 %
30	etd.aau.edu.et Internet Source	<1 %

31	humaniora.journal.ugm.ac.id Internet Source	<1 %
32	jurnal.syntaxliterate.co.id Internet Source	<1 %
33	www.tdx.cat Internet Source	<1 %
34	G. R. McGregor, D. Bamzelis. "Synoptic typing and its application to the investigation of weather air pollution relationships, Birmingham, United Kingdom", Theoretical and Applied Climatology, 1995 Publication	<1 %
35	Pawel Netzel, Tomasz Stepinski. "On Using a Clustering Approach for Global Climate Classification", Journal of Climate, 2016 Publication	<1 %
36	Yian A Chen. BMC Bioinformatics, 2004 Publication	<1 %
37	repository.usd.ac.id Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		