Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 3542-3555
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.2753

Employing cluster analysis in defining groundwater wells patterns in Rafah

Gamal R. Elkahlout[1*], Mohaned A. Elkahlout[2]
[1]School of Business Studies, Arab Open University, Riyadh, Saudi Arabia; g.elkahlout@arabou.edu.sa (G.R.E.).
[2]Faculty of Economic and Administrative Science, Al-Azhar University, Gaza; pal.mohnnad@gmail.com (M.A.E.).

Abstract: Clustering analysis techniques used in identifying homogeneity of groundwater wells patterns in terms of Chloride, Nitrate, and TDS. Data was collected through the laboratories of Palestinian Water Authority in the Rafah 2020. R-Programming is used for data analysis. Kolmogorov-Smirnov test is used to test data normality. Hierarchical clustering analysis applied to generate a cluster tree ($r>0.75$) in correlation matrix. Agglomerative Hierarchical Clustering is used to classify with "average" method and found that there are two clusters. First cluster with 28 wells and Second cluster with 9 wells. Test of homogeneity between the two clusters groups using T-test for the hierarchical clustering and found that there are significant differences in means for TDS and for Chloride between first cluster and second cluster with a significant level less than 0.01. While for Nitrates, also, there are significant differences in means between the first cluster and second cluster with a significant level less than 0.05. Different stability validation measures carried out for data concludes that for the APN and ADM measures, hierarchical clustering with two clusters again gives the best score. Three internal cluster validation measures are used. Concludes that Hierarchical clustering with two clusters performs the best in (Connective, Dunn, and Silhouette). Recall that the connectivity should be minimized, while both the Dunn Index and the Silhouette Width should be maximized. Thus, Hierarchical clustering outperforms the other clustering algorithms under each validation measure.

Keywords: *Agglomerative hierarchical clustering, Chloride nitrates, Chloride, Hierarchical clustering analysis, Hierarchical, K-means, Kolmogorov-smirnov, Nitrate, Stability, internal cluster validation measures, TDS, T-test, TDS.*

1. Introduction

This research uses clustering analysis to study patterns of water in the Rafah area, in terms of the chemical composition of drinking water wells. This pattern will be developed and compared with local and international patterns for human use of water (drinking, agricultural, industrial). Cluster analysis will depend on data clustered in groups according to similarities in chemical compounds.

Cluster analysis is one of the famous methods in multiple data analysis to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. This similarity is often determined based on specific criteria or distance metrics. There are related studies that have been performed and evaluated classifier performance of cluster analysis, and to find the characteristics that may affect the classifier performance.

A multivariate statistical technique including factor analysis, principal analysis component and cluster analysis are successfully used by Tardif (2015) to derive information from the data set about the possible influences of the environment on water quality and identify natural groupings in the set of data. These methods are important to avoid misinterpretation of environmental monitoring data due to uncertainties.

Cluster analyses are used by Wang et.al (2015) by modified varying weight. K-means cluster algorithm is proposed to classify the water quality in the Haihe River in China. The new algorithm avoids the margin of the iteration not being calculated in some cases and improves the efficiency of data processing. The result classified the seven sampling sites into four groups. The two reservoirs are the

* Correspondence: g.elkahlout@arabou.edu.sa

major drinking water headwater sites of the capital Beijing and the provincial capital of Shijiazhuang. They were relatively far from pollution sources and had better protection of water resources.

Multivariate statistical methods, studied by Muangthon (2015), namely, hierarchical agglomerative cluster analysis (HACA), discriminant analysis, principal component analysis, and the factor analysis were used to study the spatial variations of the most significant water quality variables and to determine the origin of pollution sources. Sixteen water quality parameters were initially selected and analyzed. Two spatial clusters were formed based on HACA. These clusters are designated as upper part of Nampong river and lower part of Nampong river regions.

Multivariate methods in drinking water analysis are used by Jankowska et.al (2017) during a five-year project, from 2008 to 2012, selected chemical parameters in 11 water supply networks of the Siedlce County were studied. Throughout that period drinking water was of satisfactory quality, with only iron and manganese ions exceeding the limits (21 times and 12 times, respectively). In accordance with the results of cluster analysis, all water networks were put into three groups of different water quality.

A high concentration of chlorides, sulphates, and manganese and a low concentration of copper and sodium was found in the water of Group 1 supply networks. The water in Group 2 had a high concentration of copper and sodium, and a low concentration of iron and sulphates. The water from Group 3 had a low concentration of chlorides and manganese, but a high concentration of fluorides. Using principal component analysis and cluster analysis, multivariate correlation between the studied parameters was determined, helping to put water supply networks into groups according to similar water quality.

2. The Study Problem

This study using cluster analysis in classifying groundwater wells into clusters and made comparisons of these groups with international water standards and therefore formulated the following objectives:

1. Identifying concepts of cluster analysis
2. Identifying the appropriate cluster analysis methods and algorithms and apply to study homogeneity of groundwater wells in Rafah for 2020.
3. Main findings and comparison with international water standards are extracted.

and hypothesis to be

1. There is a homogeneity between groundwater wells in the same cluster.
2. There is a homogeneity between groundwater wells in Rafah for 2020.

3. Study Methodology

The statistical method adopted in this research is cluster analysis which aims to classify the sample of observations into one or more categories based on different combinations of variables. It takes a group of cases (views) and divides them into several groups and each group contains observations that share or similar in certain qualities or characteristics. These groups are called clusters.

The purpose of this analysis is usually to discover a particular pattern that organizes views. A person can easily predict the behavior, or characteristics of other individuals or objects based on the knowledge of the groups to which they belong.

4. Concepts of Cluster Analysis

The concept of cluster analysis has been regarded as a multiple statistical method between researchers and statisticians. Unlike other statistical methods, this kind of cluster analysis does not create its own standards in terms of the existence of a single standard that can be applied to all data.

Concept of grouping in cluster analysis by dividing objects into groups so that objects in a whole and one are similar to each other, and as disparate as possible objects in other groups mass analysis is a statistical procedure for forming groups of similar objects. Scope of cluster analysis is a technique of identifying the unsupervised pattern that reveals the underlying structure or behavior of a data

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

collection without prior presumption of data, to classify system objects into categories or groups based on proximity or similarity.

5. Importance of Clustering

Cluster strategy in research can serve as a direct method for getting sorted out a tremendous information assortment with the goal that it could be better perceived, and data acquired even more rapidly. In addition, assuming the information can be truly portrayed by a predetermined number of item gatherings, the gathering marks can give a concise depiction of the information's examples of similitudes and contrasts.

Clustering strategies are used in different fields. Specialists might be focusing on strategy that serves a more crucial capacity and gives a valuable outline of the information. For instance, medication is an incredible model to discuss. Starting from characterizing infection to comprehend and treat it. Everitt et al, (2011). In industrial research, the primary function of clustering here is to perform segmentation, whether it is store, product, or client. Clients and products can be clustered into hierarchical groups based on different attributes is another example.

Another usage of clustering strategy is seen for detecting outliers like fraud transactions. Here, a cluster with all the good transactions is detected and kept as a sample. This is known as a normal cluster. Whenever something is out of the line from this cluster, it comes under the suspect section. This strategy is useful in detecting abnormal cells in the body.

Clustering is widely used to break down large datasets to create smaller data groups, which enhances the efficiency of assessing the data. Importance of clustering can be summarized in the following points:

1. Clustering helps in restarting the local search procedure and remove the inefficiency.
2. Clustering helps to determine the internal structure of the data.
3. Clustering has been used for model analysis, vector region of attraction.
4. Clustering helps in understanding the natural grouping in a dataset in order to make sense to partition the data into some group of logical groupings.
5. Clustering quality depends on the methods and the identification of hidden patterns.
6. Clustering plays a wide role in applications like marketing, economics, weblogs to identify similarity measures, image processing, and spatial research.
7. Clustering is used in outlier detections to detect credit card fraudulence

6. Clustering methods, Hierarchical Clustering

Hierarchical methods and other clustering algorithms represent an attempt to find good clusters in the data using a computationally efficient technique. It is not generally feasible to examine all possible clustering possibilities for a data set, especially a large one. In each step of the hierarchical approach: two elements are merged into a new set, so they also cannot be undone from the merge step, cannot be separated once during the procedure, and items cannot be moved to other groups, and if any errors are found they cannot be corrected.

Rough hierarchical procedure combines the two closest groups at each step and the ways of measuring the similarity or difference between the two groups should be considered. Thus, there are many different methods of measuring distance between groups that contribute to the production of new methods. (Rencher, 2002).

There are two types of hierarchical clustering methods: agglomerative and divisive. The agglomerative clustering algorithms, often called "bottom-up" methods, which start with each item being its own cluster; then, clusters are successively merged, until only a single cluster remains. While, the divisive clustering algorithms, often called "top-down" methods, do the opposite: It starts with all items as members of a single cluster; then, that cluster is split into two separate clusters, and so on for every successive cluster, until each item is in its own cluster. In more details.

6.1. Agglomerative Approaches to Generating a Hierarchical Clustering

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

The most popular agglomerative approaches of clustering are single-linkage (or nearest-neighbor), complete-linkage (or farthest-neighbor) and the cooperation between these two approaches, known as the average-linkage approach. Each of these clustering methods is defined by the way in which two clusters (which may be single items) are combined or (joined) to form a new larger cluster.
The Single-linkage approach often leads to long "chains" of clusters which are joined by singleton points near each other and a result that does not have much appeal in practice. Whereas the complete-linkage approach tends to produce many small, compact clusters. The Average-linkage approach is dependent upon the size of the clusters, but single and complete linkage approaches depend only upon the smallest or largest dissimilarity, respectively (Izenman ,2008).

6.2. Divisive Approaches to Generating a Hierarchical Clustering

Divisive algorithms have two approaches, monothetic and polythetic. In a monothetic approach, the division of a group into two subgroups is based on a single variable, whereas the polythetic approach uses all p variables to make the split. If the variables are binary (quantitative variables can be converted to binary variables), the monothetic approach can easily be applied. The separating of variables into two groups is based on presence or absence of an attribute. The variable (attribute) is chosen based on a chi-square statistic or an information statistic.

In a polythetic approach in each step, the items are divided into "splinter" group (say, cluster A) and "remainder" (cluster B). The splinter group is initiated by extracting that item which has the largest average dissimilarity from all other items in the data set; this item is set up as cluster. Given this separation of the data into A and B, we calculate for each item in cluster B, the following two quantities:

(1) the average dissimilarity between that item and all other items in cluster B, and;

(2) the average dissimilarity between that item and all items in cluster A.

Then, we calculate the difference between (1) and (2) for each item in B. If all differences are negative, we stop the algorithm. Comparatively, if any of these differences are positive (indicating that the item in B is closer on average to cluster A than to the other items in cluster B), we take the item in B with the largest positive difference, then we shifted it to A, and repeat the procedure.

7. Clustering methods, Nonhierarchical Methods

Three nonhierarchical techniques will be discussed: partitioning, *k*-Means, and *k*-medoids. Among these three methods, partitioning is mostly used.

7.1. Partitioning

In the partitioning approach, observations are separated into *g* clusters without using a hierarchical approach based on a matrix of distances or similarities between all pairs of points. This method, sometimes called optimization method using the following steps:

1. Select an initial partition with *k* clusters.
2. Generate a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster centers as the centroids of the clusters.
4. Repeat steps 2 and 3 until an optimum value of the criterion function is found.
5. Adjust the number of clusters by merging and splitting existing clusters or by removing small, or outlier clusters.

Repeat steps 2 through 5 until the cluster membership stabilizes.

7.2. K-Means

This approach allows the items to be moved from one cluster to another. Henceforth, the researcher will continue using the notation *g* rather than *k* for the numbers of clusters. We first select *g* items to serve as *seeds*. These are later replaced by the centroids (mean vectors) of the clusters.

Choose the first *g* points in the data set (again subject to a minimum distance requirement), select the *g* points that are mutually farthest apart, find the *g* points of maximum density, or specify *g* regularly spaced points in a grid like pattern (these would not be actual data points). Rencher (2002).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

After all items are assigned to clusters, each item is examined to see if it is closer to the centroid of another cluster than to the centroid of its own cluster. If so, the item is moved to the new cluster and the two cluster centroids are updated. This process is continued until no further improvement is possible.

7.3. K-medoids Clustering

The *k*-medoids algorithm attempts to minimize squared error, and the distance between points labeled to be in a cluster and a point designated as the center of that cluster. *K*-medoids is a partitioning technique of clustering that groups the data set of *n* objects into *k* clusters with *k* known a priori. A useful tool for determining *k* is the Silhouette. It could be more robust to noise and outliers as compared to *k*-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared "Euclidean" distances. The possible choice of the dissimilarity function is very rich, but in this research, we used the squared Euclidean distance.

A *k*-medoids of a finite dataset is a data point from this certain set, whose average dissimilarity to all the data points for it is minimal. In other words, it is the most centrally located point in the set. According to Mirkes (2011), the most common recognition of *k*-medoid clustering is the partitioning around Medoids (PAM) algorithm and it works as follows:

1- Initialize a randomly select *k* of the *n* data points as the medoids, then associate each data point to the closest medoid and for each medoid m and each associated data point.

2- "m" swap "m and o" and compute the total cost of the configuration (that is, the average dissimilarity of o to all the data points associated to m).

3- select the medoid o with the lowest cost of the configuration.

4- At last, repeat alternating steps 2 and 3 until there is no change in the assignments.

Differences between hierarchical clustering and non-hierarchical clustering are given in Tabel(1).

Table 1.
Differences between hierarchical clustering and non-hierarchical clustering.

No	Hierarchical clustering (HC)	Non hierarchical clustering
1	It involves creating clusters in a predefined order from top to bottom	It involves formation of new clusters by merging or splitting clusters instead of following HC order.
2	It is considered less reliable than non-HC	It is comparatively more reliable than HC
3	It is considered slower than non-HC	It is comparatively faster than HC
4	Avoid applying this technique when data with high level of errors	It can work better then HC even when error is there.
5	It is comparatively easier to read and understand.	Clusters are difficult to read and understand as compared to HC
6	It is relatively unstable than non HC	It is a relatively stable technique.

Dou to differences arise in Tabel (1), HC offers distinct advantages and limitations compared to non-hierarchical clustering methods in real-world applications. One of the primary benefits of HC is its ability to create a dendrogram, which allows users to visualize and select an appropriate number of clusters (*k*) based on the data's inherent structure. This flexibility is particularly useful in domains where cluster hierarchies naturally exist, as it does not require prior knowledge of the number of classes present in the dataset. HC is applied in analyzing this research data.

8. Data Analysis and Presentation

The data was collected through the laboratories of the Palestinian Water Authority in the Rafah for year 2020. It is subjected to chemical analysis and registration of several chemical compounds. This study concern with three chemical components: Chloride, Nitrate, and Total dissolved solids (TDS).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

Chloride exists in all natural waters with widely varying concentrations reaching a maximum in seawater up to 35,000 mg/l. In fresh waters, the sources include soil and rock formations, sea spray, and waste discharges. Sewage contains large amounts of chloride, as do some industrial effluents as shown in Table 2.

Table 2.
Classification of water based (Chloride) in mg/L.

Water Class / Using	Unit	Guideline before transform	Guideline after transform
Accepted/Human, agricultural	mg/L	Less than 300	Less than 2.477
Not accepted	mg/L	Over than 300	Over than 2.477

Source: WHO (1996).

Nitrates are naturally occurring ions (NO_3) that are part of the nitrogen cycle. This ion is stable form of combined nitrogen for oxygenated systems, chemically unreactive, and can be reduced by microbial action. The nitrite ion (NO_2) contains nitrogen in a relatively unstable oxidation state. Chemical and biological processes can further reduce nitrite to various compounds or oxidize it to Nitrate as shown in Table 3.

Table 3.
Classification of water based (Nitrate) in mg/L.

Water class / Using	Unit	Guideline before transform	Guideline after transform
Accept /Human	mg/L	Less than 50	Less than 1.698
Accepted / Agricultural	mg/L	50 -100	1.698 -2
Not Accepted	mg/L	Over 100	Over 2

Source: WHO (2007).

Total dissolved solids (TDS) used to describe the inorganic salts and small amounts of organic matter present in solution in water. The principal constituents are usually calcium, magnesium, sodium, and potassium cations and carbonate, chloride, sulfate, and Nitrate anions as shown in Table 4.

Table 4.
Classification of water based on (TDS) in mg/L.

Water Class / Using	Unit	Guideline before transform	Guideline after transform
Accepted/Human	mg/L	Less than 1000	Less than 3
Accepted/Agricultural	mg/L	1000 – 3000	3-3.477121
Not Accepted	mg/L	Over 3000	Over 3.4771

Source: WHO (2007).

To perform our data analysis, many statistical procedures assume that the variables are normally distributed. A significant violation of the assumption of normality can seriously increase the chances of the researcher committing either a Type I or II error (depending on the nature of the analysis and the non-normality). However, Micceri (1989) points out that true normality is exceedingly rare in education and psychology. Thus, one reason (although not the only reason) researchers utilize data transformations is to improve the normality of variables.

There are multiple options for dealing with non-normal data. First, the researcher must make certain that the non-normality is due to a valid reason (real observed data points). Invalid reasons for non-normality includes mistakes in data entry or missing data values not declared missing.

However, not all non-normality is due to data entry error or non-declared missing values. Two other reasons for non-normality are the presence of outliers (scores that are extreme relative to the rest of the sample) and the nature of the variable itself. There is great debate in science about whether outliers should be removed or not.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

The normality test for the Water wells data in Rafah (2020) was conducted using the Kolmogorov-Smirnov test, revealing a p-value greater than 0.05 after data transformation, indicating that the data conforms to a normal distribution. Consequently, this allows for the application of parametric tests, which assume the data follows a specific distribution.

Table 5.
Test of normality (Kolmogorov-Smirnov).

Process	Variables	Kolmogorov-Smirnov Z	P-value	Decision
Before	TDS	1.468	0.027	Not Normal
	Chloride	1.771	0.004	Not Normal
	Nitrate	0.962	0.313	Normal
After	TDS	0.672	0.757	Normal
	Chloride	0.704	0.704	Normal
	Nitrate	0.953	0.324	Normal

9. Hierarchical Clustering Analysis

R programming language is used to perform hierarchical clustering and generate a cluster tree using " *hclust*()" function. The function "*cutree()*" can be used to cut a tree which reflects data is to compute the correlation between the *cophenetic* and the original distance data generated by "*dist*()" function.

If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the original distance matrix**.** The closer the value of the correlation is to 1, the more accurately the clustering solution reflects your data. Values above 0.75 are felt to be good. Results are summarized in Table 6.

Table 6.
Optimal scores of agglomerative clustering and cophenetic of clusters.

Hierarchical	Cophenetic	AC
Ward	0.9049138	0.9361213
Single	0.889809	0.9265208
Complete	0.9266485	0.941376
average	0.9361161	0.9361213

It is concluded that the best method was "**average**", where the Cophenetic is almost equal to Agglomerative Clustering (AC) with score equal to 0. ***9361213***

Using Agglomerative Hierarchical Clustering to classify and using the "average" to be clustering group. This method is based when two clusters are be joined, the distance of the new cluster to any other cluster is calculated as the average of the distances of the soon to be joined clusters to that other cluster in the data. Using the Hierarchical Clustering (HC) to find the K-cluster in Rafah as given in Figure 1.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

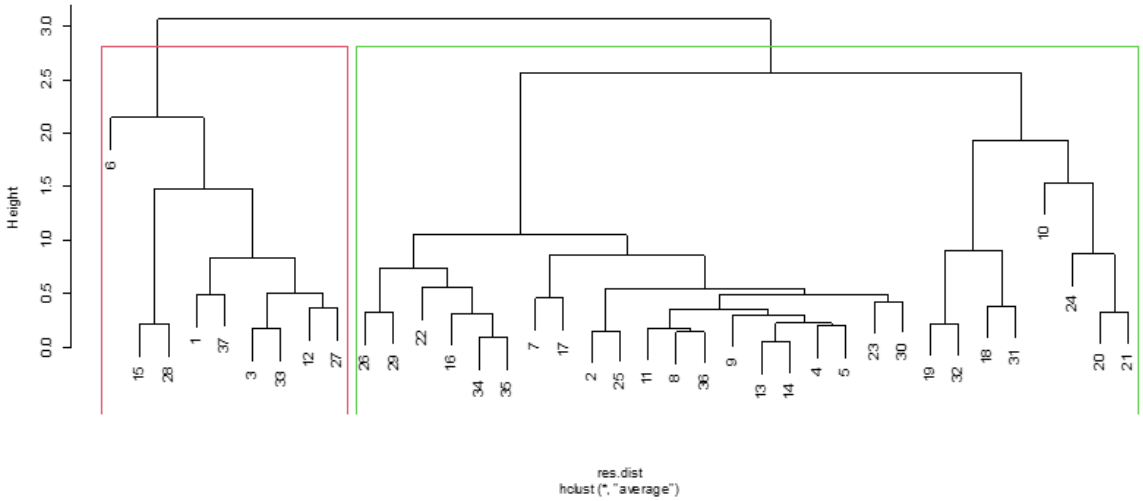

Figure 1.
Average clustering among groundwater wells in Rafah.

Hierarchical analysis was applied using "average" method on groundwater wells in Rafah Governorate for the year (2020), where we found two clusters and summarized in Table 7. where we conclude that for the first cluster there are 28 wells (P/163, AlShoka, P/147, P/170, P/15, P/138, P/144A, P/145, AlSika, AlSafa, P/174, AlShaout, W02, W03, IbnTaimiya, UNDP, Muraj, AlZohor, P/164, P/138, P/144A, P/173, P/174, AlShaout, W01, Muraj, P/148, and P/148) had the same (Nitrate Guideline) which was (Not Acceptability for human and agricultural using) And the same (T.D.S Guideline) which was (Not Acceptability for human and agricultural using), and the same (Chloride Guideline) which was (Not Acceptability for human and agricultural using).

While for the second cluster, 9 wells (P/10, IbnTaimiya, Tika, P/172, P/173, P/153, Malaysian, AlNaser2, and P/169) had the same (Nitrate Guideline) which was (Not Acceptability for human and agricultural using) And the same (T.D.S Guideline) which was (Not Acceptability for human and agricultural using), and the same (Chloride Guideline) which was (Not Acceptability for human and agricultural using).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

Table 7.
K- cluster of Groundwater wells in Rafah by average method (2020).

No.	Wells	Well name	Well no.	k-cluster	Nitrate guideline	T.D.S guideline	Chloride guideline	Nitrate	Chloride	T.D.S
1.	2	Al Naser 1	P/163	1	Not accept	Not accept	Not accept	5.09	6.84	7.78
2.	4	Baladia Al Shoka	Al Shoka	1	Not accept	Not accept	Not accept	5.41	7.05	8.09
3.	5	Airport	P/147	1	Not accept	Not accept	Not accept	5.37	7.02	7.92
4.	7	Al Shoka Al Jamia	P/170	1	Not accept	Not accept	Not accept	5.84	6.68	7.74
5.	8	Zourob	P/15	1	Not accept	Not accept	Not accept	5.47	6.92	8.09
6.	9	Abu Zuhri	P/138	1	Not accept	Not accept	Not accept	5.32	6.77	7.87
7.	10	Canada New	P/144A	1	Not accept	Not accept	Not accept	4.44	6.46	7.34
8.	11	Al Hashash	P/145	1	Not accept	Not accept	Not accept	5.55	7.03	7.99
9.	13	Al Sika	Al Sika	1	Not accept	Not accept	Not accept	5.26	6.98	8.04
10.	14	Al Safa	Al Safa	1	Not accept	Not accept	Not accept	5.26	6.98	7.99
11.	16	TRC3	P/174	1	Not accept	Not accept	Not accept	5.42	6.19	7.19
12.	17	Al Shaout	Al Shaout	1	Not accept	Not accept	Not accept	5.82	6.93	8.09
13.	18	W02 /Al Jadied 2	W02	1	Not accept	Not accept	Not accept	4.09	8.87	9.39
14.	19	W03 /Al Jadied 3	W03	1	Not accept	Not accept	Not accept	3.99	8.29	8.91
15.	20	Ibn Taimiya	Ibn	1	Not accept	Not accept	Not accept	4.44	7.39	8.39
16.	21	UNDP Housing	UNDP	1	Not accept	Not accept	Not accept	4.29	7.58	8.25
17.	22	Muraj	Muraj	1	Not accept	Not accept	Not accept	5.42	6.66	7.59
18.	23	Al Zohor	Al Zohor	1	Not accept	Not accept	Not accept	5.16	7.21	8.19
19.	24	UNRWA	P/164	1	Not accept	Not accept	Not accept	3.83	7.36	8.18
20.	25	Abu Zuhri	P/138	1	Not accept	Not accept	Not accept	5.14	6.87	7.90
21.	26	Canada New	P/144A	1	Not accept	Not accept	Not accept	4.97	6.46	7.37
22.	29	TRC2	P/173	1	Not accept	Not accept	Not accept	5.17	6.37	7.36
23.	30	TRC3	P/174	1	Not accept	Not accept	Not accept	5.43	7.23	8.28

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

No.	Wells	Well name	Well no.	k-cluster	Nitrate guideline	T.D.S guideline	Chloride guideline	Nitrate	Chloride	T.D.S
24.	31	Al Shaout	AlShaout	1	Not accept	Not accept	Not accept	4.32	8.91	9.52
25.	32	W01 /Al Jadied 1	W01	1	Not accept	Not accept	Not accept	4.09	8.18	9.00
26.	34	Muraj	Muraj	1	Not accept	Not accept	Not accept	5.56	6.32	7.32
27.	35	TalAlSultan/PWA	P/148	1	Not accept	Not accept	Not accept	5.61	6.34	7.27
28.	36	TalAlSultan/PWA	P/148	1	Not accept	Not accept	Not accept	5.55	6.87	8.05
1.	37	UNRWA Rafah	P/10	2	Not accept	Not accept	Not accept	4.62	4.61	6.11
2.	33	Ibn Taimiya	IbnTaimiya	2	Not accept	Not accept	Not accept	4.46	5.45	6.67
3.	27	Tika	Tika	2	Not accept	Not accept	Not accept	4.62	5.14	6.27
4.	28	TRC1	P/172	2	Not accept	Not accept	Not accept	3.69	5.71	6.76
5.	15	TRC2	P/173	2	Not accept	Not accept	Not accept	3.71	5.57	6.61
6.	12	Al Eskan	P/153	2	Not accept	Not accept	Not accept	4.68	5.44	6.46
7.	6	Malaysian	Malaysian	2	Not accept	Not accept	Not accept	5.56	4.47	6.57
8.	3	Al Naser 2	Al Naser 2	2	Not accept	Not accept	Not accept	4.40	5.33	6.58
9.	1	Al Naser 3	P/169	2	Not accept	Not accept	Not accept	4.33	4.77	6.08

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

Therefore, a test of homogeneity between the first cluster group and the second cluster group has been done. Results are given in Table 8 for the three chemical compounds and conclude that

- There are significant differences in means for TDS between the first cluster and second cluster with a significant level less than 0.01 and t- test = 7.679.
- There are significant differences in means for Chloride between the first cluster and second cluster with a significant level less than 0.01 and t- test = 7.682.
- There are significant differences in means for Nitrate between the first cluster and second cluster with a significant level less than 0.05 and t- test = 2.629.

Table 8.
T-test for chemical compounds in Rafah (2020).

		N	Mean	Std. deviation	T	Sig.
T.D.S	1.00	28	8.0391	0.59704	7.679	0.000
	2.00	9	6.4577	0.24615		
Chloride	1.00	28	7.0984	0.70717	7.682	0.000
	2.00	9	5.1644	0.44780		
Nitrate	1.00	28	5.0471	0.59694	2.629	0.013
	2.00	9	4.4545	0.55760		

10. Stability Validation measures

The stability measures compare the results from clustering based on the full data to clustering based on removing each column, one at a time. These measures work especially well if the data are highly correlated, which is often the case in high-throughput genomics data. There are four measures given by Brock et al. (2008), Datta (2006) and Yeung et al. (2001). In all cases the average is taken over all the deleted columns, and all measures should be minimized.

10.1. Average proportion of non-overlap (APN)

The NP which measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed.

10.2. Average Distance (AD)

Th AD measure computes the average distance between observations placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. AD has a value between zero and ∞, and smaller values are preferred

10.3. Average Distance between Means (ADM)

The ADM measure computes the average distance between cluster centers for observations placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. Currently, ADM only uses the Euclidean distance. It also has a value between zero and ∞, and again smaller values are preferred.

10.4. Figure of Merit (FOM)

The FOM measures the average intra-cluster variance of the observations in the deleted column, where the clustering is based on the remaining (undeleted) samples. This estimates the mean error using predictions based on the cluster averages. The final score is averaged over all the removed columns, and has a value between zero and ∞, with smaller values equaling better performance. Stability measures include the APN, AD, ADM, and FOM. The measures should be minimized in each case. Stability validation requires more time than internal validation, since clustering needs to be redone for each of the datasets with a single column removed. Results of stability validation measures are given in Table 9 and concludes that for the APN and ADM measures, hierarchical clustering with two clusters again gives the best score. For the other measures, hierarchical with two clusters has the best score.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

Table 9.
Optimal scores for stability validate measure in Rafah.

Validation measures	Score	Method	Clusters
APN	0.09459459	Hierarchical	2
AD	1.62050621	Hierarchical	2
ADM	0.46880538	Hierarchical	2
FOM	0.76602998	Hierarchical	2

11. Cluster Validation

To validate the use of cluster analysis, the hypothesis of elements or groups in the sample from which the sample is available should be tested. For example, the hypothesis could be representation of the population of a single distribution such as multivariate distribution, or the observations originated from a uniform distribution.

Also, cross-checking policies can be used to validate or aggregate the aggregation results. The data is randomly divided into two sub-groups. To illustrate, the block analysis of A and B is performed separately on both A and B. The results should be similar if the groups are valid.

According to Gordon (1999) and Milligan (1996) an alternative approach for cluster validating can be done as the following:

1. Use some clustering method to partition sub-set of "A" into "g" clusters.
2. Partition subset "B" into "g" clusters in two ways:

(a) Assign each item in "B" into the cluster in "A", which is closest to it using, for example, the distance to cluster centroids.

(b) Use the same clustering method on "B" that was used on "A."

3. Compare the results of (a) and (b) in step 2.

Many different cluster validity methods have been proposed without any a priori class information. According to (Chow et al., 2002) and (Chow et al,. 2004) the clustering validation is a technique to find a set of clusters that best fits natural partitions (number of clusters) without any class information.

Furthermore, the analysis techniques of cluster can be based on external validation based on prior knowledge about data follows, and internal validation based on the information secured inside in the data alone. Halkidi et al. (2001).

If these two types of validation's are blocked to determine the appropriate number of groups from a data set, the consideration will follow certain options. One option is to use external validation indexes that require prior knowledge about the data used, but it is difficult to determine whether they can be used in data that has real problems. Another option is to use internal validity indexes that do not require advance information from the dataset.

Here we used the internal validation measures are the connectivity, Silhouette Width, and Dunn Index. The neighborhood size for the connectivity is set to 10 by default. Results are summarized in Table 10 and Table 11.

Table 10.
Compare between internal validation measures in Rafah (2020).

Clustering method	Validation Measures	(Score) Number of clusters=2	(Score) Number of clusters=3	(Score) Number of clusters=4
Hierarchical	Connectivity	4.0353	10.1048	12.7516
	Dunn	0.2953	0.2414	0.2414
	Silhouette	0.5023	0.5813	0.5481
K-means	Connectivity	4.0353	8.9873	11.6341
	Dunn	0.2953	0.2593	0.3641
	Silhouette	0.5023	0.5934	0.5584

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753

Table 11.
Optimal Scores for internal validation measures in Rafah (2020).

Internal validation	Score	Method	Clusters
Connectivity	4.0353	Hierarchical	2
Dunn	0.2953	Hierarchical	2
Silhouette	0.5023	Hierarchical	2

Concludes that Hierarchical clustering with two clusters performs the best in (Connective, Dunn, Silhouette) Recall that the connectivity should be minimized, while both the Dunn Index and the Silhouette Width should be maximized. Thus, it appears that Hierarchical clustering outperforms the other clustering algorithms under each validation measure.

12. Conclusions

Water wells data in Rafah (2020) are normally distributed due to Kolmogorov-Smirnov test. Consequently, using T test in independent. Hierarchical Clustering Analysis involves computing the correlation between the cophenetic distance and the original distance data to validate the clustering solution. A strong correlation indicates that the cluster tree accurately reflects the data, with values above 0.75 considered satisfactory.

Optimal Scores of Agglomerative Hierarchical Clustering (AHC) and Cophenetic of clusters reveal that the best method was Cophenetic which is almost equal to Agglomerative Clustering (AC) with a high score around 0.94. By using (AHC) to find the K-cluster applying "average" method on groundwater wells in Rafah for 2020. Two clusters arise with 28 wells in one cluster and 9 wells in the other. With the same Nitrate, TDS and Chloride guidelines. All those guidelines are not acceptable levels for human and agricultural use.

A test of homogeneity between the two-cluster group has been done by using T-test for the three chemical compounds and conclude that there are significant differences in means with a significant level less than 0.01 for TDS and for Chloride, and less than 0.05 for Nitrates.

Stability validation measures conclude that for the APN and ADM measures, hierarchical clustering with two clusters gives the best score.

In clusters validations we used internal validation measures, which are Connectivity, Dunn Index, and Silhouette Width with a neighborhood size for the connectivity set to 10 and found that Hierarchical clustering with two clusters performs the best in (Connective, Dunn, Silhouette).

Acknowledgement:

The authors extend their appreciation to the Arab Open University for funding this work through AOU research fund No. (AOURG-2023-016).

References

[1] Brock G., Pihur V., and Datta S. (2008): clValid: an R package for cluster validation, Journal of Statistical Software,25,4.

[2] Chou, C.H, Su M.C and Lai, E. (2002): Symmetry as A new measure for Cluster Validity. 2 th. WSEAS Int.Conf. scientific Computation and Soft Computing, Crete, Greece, pp. 209-213.

[3] Chow C.H, Su M.C and Lai E. (2004): A new Validity Measure for Clusters with Different Densities. Pattern Anal. Applications, 7, pp.2005-2020.

[4] Datta S. (2006): Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, BMC Bioinformatics, 7:397.

[5] Everitt S. B.& Landau S., Leese M. , Stahl D. , (2011): Cluster Analysis, 5th Edition, Willey series in probability and statistics, King's College London, UK

[6] Gordon A. D.(1999): Classification. Chapman & Hall/CRC, Boca Raton, FL, 2 edition,

[7] Halkidi M., Batistakis Y.,Vazirgiannis M. (2001): On Clustering Validation Techniques. Department of Informatics, Athens University of Economics and Business.
[8] Izenman, Alan Julian, (2008): Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning, Springer.
[9] Jankowska, j , Radzka, E. and Katarzyna, R. (2017): Principal component analysis and cluster analysis Multivariate assessment of water quality, journal of Ecological Engineering, 18, 2.
[10] Muangthong, S. (2015): Assessment of surface water quality using multivariate statistical techniques: A case study of the Nampong River Basin, Thailand, The Journal of Industrial Technology, Vol. 11
[11] Mirkes E.M. (2011): K-means and K-medoids applet. University of Leicester, http://www.math.le.ac.uk
[12] Micceri, T (1989): The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105,1, 156-166.
[13] Milligan G. W. (1996): Clustering validation: results and implications for applied analyses. Clustering and Classification, 341-375.
[14] Rencher C. A. (2002): Methods of Multivariate Analysis, Brigham Young University, Canda.
[15] Tardif, Geneviève (2015): Multivariate Analysis of Canadian Water Quality Data, PhD thesis, University of Ottawa,177
[16] Wang, X. , Zou, Z. Zou, H.(2015): An Enhanced K-Means Algorithm for Water Quality Analysis of The Haihe River in China, Int. J. of Environmental Research and Public Health 12(11):14400-14413
[17] World Health Organization (1996): WHO handbook on Indoor Radon: A public health perspective. Geneva.
[18] World Health Organization (2007): Guidelines for drinking-water quality, Second edition.
[19] World Health Organization, (2008): Guidelines for drinking-water quality, Third edition.
[20] Yeung K.Y., Haynor D.R., and Ruzzo W.L.(2001): Validating clustering for gene expression data, Bioinformatics, 17(4):309-18

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 3542-3555, 2024
DOI: 10.55214/25768484.v8i6.2753