International Journal of Analysis and Applications

Impact of Dataset Scaling on Hierarchical Clustering: A Comparative Analysis of Distance-Based and Ratio-Based Methods

Ali Rashash R. Alzahrani*

Mathematics Department, Faculty of Sciences, Umm Al- Qura University, Makkah, Saudi Arabia
**Corresponding author:* arrzahrani@uqu.edu.sa

ABSTRACT. In this study, the distance-based agglomerative hierarchical clustering techniques were compared to a ratio-based approach. Two real datasets, which were also used in a prior study by Roux (2018), were considered. Firstly, it was observed that the type of scaling applied to the datasets was found to affect the results of hierarchical clustering. Thus, various scaling methods were employed prior to implementing hierarchical clustering. Furthermore, two rank-based goodness-of-fit measures were used to evaluate the hierarchical clustering methods. In contrast to Roux (2018) findings, it was observed that the distance-based methods, such as Median linkage, Average linkage, and centroid linkage, performed better than the ratio-based method. The ratio-based methods also showed issues with branch crossing in the hierarchical clustering dendrogram. Consequently, this study illustrates that, with appropriate dataset scaling, the distance-based methods outperform ratio-based methods in terms of goodness-of-fit measures.

1. INTRODUCTION

Agglomerative hierarchical clustering is a widely employed technique in data analysis and machine learning for grouping similar data points into clusters in a hierarchical manner. This method begins with each data point as a separate cluster and iteratively merges clusters based on a chosen linkage criterion until a single cluster encompassing all data points is formed [1].

To link the object together in a clustering algorithm, proximity measures are employed. The proximity measure (similarity/dissimilarity) are calculated using features or parameters in the

Received Nov. 1, 2023

2020 Mathematics Subject Classification. 14M06.

Key words and phrases. distance type methods; ratio type method; median linkage; centroid linkage; average linkage.

https://doi.org/10.28924/2291-8639-22-2024-36

ISSN: 2291-8639

dataset. To get the dissimilarity between two cases in the study (case *i* and *j*), knowing that each case includes *p* parameters $\{(x_{i1}, x_{i2}, \dots, x_{ip})$ and $(x_{j1}, x_{j2}, \dots, x_{jp})\}$, the dissimilarity measure can be considered as the difference between these p parameters in case *i* and *j*. One commonly used dissimilarity measure is Euclidean distance, which is the root of square differences between *p* parameters. The Euclidean distance [2] between them can be calculated as:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \quad (1)$$

The Euclidean distance is a measure of dissimilarity which can be calculated for distance between single cases. In agglomerative hierarchical clustering, only in the first step all cases are located in a cluster with a single case [2].

The closest or the most similar cases are merged in each step of agglomerative hierarchical clustering. The distance matrix for all pairs of observations is calculated. The two cases which have the least pairwise distance is merged. In the next step one of the clusters includes more than 1 case. The dissimilarity of other clusters with this cluster requires a linkage [3]. Commonly used linkage methods are average linkage, single linkage, complete linkage method [3]. Assume case 1 and 2 are merged in the first step and now they are in one cluster the distance between them and case 3 can be calculated by using average linkage $AL[(1,2),(4)]$ given by:

$$AL[(1,2),(4)] = \frac{d(x_1,x_4)+d(x_2,x_4)}{2} \quad (2)$$

where *d* is the Euclidean distance. In the next step if the object formed by observations {1,2, 4} has the least average linkage compared with other objects, then the average linkage is calculated as:

$$AL((1,2,4),(3)) = \frac{d(x_1,x_3)+d(x_2,x_3)+d(x_4,x_3)}{3} \quad (3)$$

If cluster 1 includes observations {1 , 2} and cluster 2 includes {3, 4} then the average linkage will be:

$$AL((1,2),(3,4)) = \frac{d(x_1,x_3)+d(x_1,x_4)+d(x_2,x_3)+d(x_2,x_4)}{4} \quad (4)$$

On the other hand, single linkage takes the minimum distance of observations in cluster 1 and observations in cluster 2. The complete linkage takes the maximum distance between cases in cluster 1 and cluster 2 and considers it as the node level. In agglomerative hierarchical clustering objects are merged step by step until all the cases are located in one cluster.

To evaluate the structure of the dendrogram in the hierarchical clustering algorithm, all the pairs of objects in terms of their Euclidean distance and the ultrametric distance which is

derived from their node level in the dendrogram are considered. Kendall's tau [4], Goodman-Kruskal's coefficients [5] are rank based measures which are used as goodness of fit measure for hierarchical clustering, and both are useful for evaluating the hierarchical clustering structure. Since GK considers only comparable pairs, it is more appropriate to be used as a goodness of fit measure compared with KT which considers all possible pairs including ties and non-comparable pairs.

Thus, the primary objective of this study is to compare various hierarchical clustering methods while taking into account the influence of different dataset scaling techniques. Our observations indicate that the choice of scaling method significantly affects the structure of hierarchical clustering. The secondary goal of this study is to compare the outcomes of different hierarchical clustering methods with those obtained in a recent study conducted by Roux [6]. Additionally, is to reevaluate the conclusions drawn in Roux's study in light of the findings from our own investigation. To ensure a fair comparison, the same datasets was employed as those utilized by Roux [6].

2. SCALING METHODS

The dataset can be normalized or standardized by several methods. In this study the real datasets used by scaling them using the following scaling techniques:

i. **Mean Absolute Deviation from the Median**: This scaling method involves calculating the absolute difference between each data point and the median of the dataset, then finding the average of these absolute differences. It measures the average dispersion of data points around the median. Suppose y_i is the scaled data point, x_i is the original data points and $\hat{x}$ is the median of x_i, the scaled mean absolute deviation from the median can be computed using:

$$y_i = n^{-1} \sum_{i=1}^{n} |x_i - \hat{x}|.$$

ii. **Median Absolute Deviation**: This scaling method involves finding the median of the absolute differences between each data point and the median of the dataset. It quantifies the spread of data points from the median while being robust to outliers. Suppose y_i is the scaled data point, x_i is the original data points and $\hat{x}$ is the median of x_i, the scaled median absolute deviation can be computed using:

$$y_i = \underset{0.5}{\arg}\left(\sum_{i=1}^{n}|x_i - \hat{x}|\right).$$

iii. **Interquartile Range**: The interquartile range (IQR) is a scaling method that measures the spread of data by calculating the difference between the third quartile (Q3) and the first quartile (Q1) in a dataset. It is a measure of the middle 50% of the data's distribution and is also robust to outliers. Suppose y_i is the scaled data point, x_i is the original data points and $\hat{x}$ is the median of x_i, the scaled interquartile range can be computed using:

$$y_i = IQR^{-1}\sum_{i=1}^{n}(x_i - \hat{x}).$$

iv. **Standard Deviation**: The standard deviation is a widely used scaling method that measures the average deviation of data points from the mean (average) of the dataset. It provides a comprehensive assessment of data dispersion, but it can be sensitive to outliers. Suppose y_i is the scaled data point, x_i is the original data points and $\hat{x}$ is the median of x_i, the scaled interquartile range can be computed using:

$$y_i = \left(\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\right)^{-1}\sum_{i=1}^{n}(x_i - \hat{x}).$$

v. **No Scaling**: In this case, "no scaling" means that the data is used as is, without any specific scaling applied. This can be useful when the data is already on a compatible scale or when scaling isn't considered necessary for the analysis. Suppose y_i is the scaled data point, x_i is the original data points the no scaled can be computed using:

$$y_i = x_i.$$

These scaling were used for various types of hierarchical clustering algorithms. To see the influence of various types of scaling on the structure of the hierarchical clustering result.

3. MATERIAL AND METHODS

3.1. Real datasets:

Two real datasets were used in this study which are the same as the datasets which were used by Roux [6].

Pottery dataset:

The chemical dataset of Romano-Bristish pottery [7]. This dataset includes 48 cases. Three of these are unusable, thus they are removed from the data. Therefore, the analysis data consists of 45

cases with 9 quantitative variables. The data were first standardized by various types of scaling and then used in hierarchical clustering.

Fisher's Irish dataset:

This dataset is a widely used dataset in various studies of statistical analysis [8]. It includes 150 cases and four parameters of Sepal Length, Sepal Width, Petal Length and Petal Width. This data is also standardized with various scaling methods and used in hierarchical clustering.

3.2. Hierarchical clustering

In hierarchical clustering, the data points are initially treated as individual clusters, and then they are successively merged or divided based on their similarity or dissimilarity to form a tree-like structure known as a dendrogram. In this study two forms of agglomerative hierarchical clustering were considered.

Hierarchical clustering with distance metrics:

Various linkage methods are employed for merging the objects together in hierarchical clustering. In this study Average linkage, single linkage, complete linkage, centroid and Median method have been used as metric for merging the observations.

i. **Average linkage**: Average linkage clustering, also known as UPGMA (Unweighted Pair Group Method with Arithmetic Mean), is a hierarchical agglomerative clustering method used in data analysis. It is primarily employed for grouping data points into clusters based on their similarity or dissimilarity. The average linkage clustering is computed using:

$$AL(C_1, C_2) = \frac{1}{n_1 n_2} \sum \{d(x_i, x_j) : i \in C_1, j \in C_2\}.$$

ii. **Single Linkage**: Single linkage clustering, also known as single-link clustering or nearest-neighbor clustering, is a hierarchical agglomerative clustering method used in data analysis and data mining. In single linkage clustering, data points or objects are initially treated as individual clusters, and at each step of the clustering process, the two closest clusters are merged into a single cluster. The single linkage clustering is computed using:

$$SL(C_1, C_2) = Min\{d(x_i, x_j) : i \in C_1, j \in C_2\}.$$

iii. **Complete Linkage**: Complete linkage clustering is a hierarchical agglomerative clustering method used in data analysis and machine learning. It is a bottom-up approach where data points are initially treated as individual clusters and are successively merged into

larger clusters based on their pairwise dissimilarity or distance. The complete linkage is computed using:

$$CL(C_1, C_2) = Max\{d(x_i, x_j): i \in C_1, j \in C_2\}.$$

iv. **Median linkage**: Median linkage clustering, also known as the median method or UPGMM (Unweighted Pair Group Method with Median), is a hierarchical clustering technique used in data analysis and data mining. It is commonly applied to group similar data points or objects into clusters based on their pairwise dissimilarity or distance measures when extreme values are suspected. The median linkage is computed using:

$$ML(C_1, C_2) = Median\{d(x_i, x_j): i \in C_1, j \in C_2\}.$$

For each of the methods in i - iv, the *d* is Euclidean distance, C_1 and C_2 are two clusters, n_1 is the number of cases in C_1 and n_2 is the number of cases in C_2.

v. **Centroid method**: Unlike the four methods above, the centroid method is derived from distance which is the distance between the centroid of objects in one cluster with the centroid of objects in another cluster.

$$Centroid\ distance(C_1, C_2) = d(\bar{x}_1, \bar{x}_2)$$

where *d* is still the Euclidean distance, C_1 and C_2 are two clusters, $\bar{x}_1$ is the centroid of objects in C_1 and $\bar{x}_2$ is centroid of objects in C_2. The centroid mean or centers are computed using:

$$\bar{x}_c = \frac{1}{n_c}\sum\{x_i : c \in C\}.$$

Hierarchical clustering with ratio-type metrics:

This study does not primarily focus on methods that involve ratios. To make meaningful comparisons, a specific type of hierarchical clustering method was only, known as relative hierarchical clustering, as presented by Mollineda & Vidal [9]. This method, which is classified as a ratio-based approach, was identified as the top performer for agglomerative hierarchical clustering in both the real datasets (Pottery and Irish) studied by Roux [6]. In the method introduced by Mollineda & Vidal [9], it's important to note that it doesn't just take into account the dissimilarity between clusters when merging objects. It also factors in the distances between the clusters and the other clusters in the denominator of the dissimilarity measure. This relative distance metric calculates dissimilarity by considering the distance between two objects divided

by the minimum of the average distances of each object with the other clusters. This calculation is referred to as the isolation function.

$$\gamma(i,j) = \frac{\sum\{\mathrm{d(i,k)} \mid \mathrm{k} \in \mathrm{C}, \mathrm{k} \neq \mathrm{j}\}}{NC-2}$$

$$\gamma(j,i) = \frac{\sum\{\mathrm{d(j,k)} \mid \mathrm{k} \in \mathrm{C}, \mathrm{k} \neq \mathrm{i}\}}{NC-2}$$

where C is the clusters at the current step of hierarchical clustering and NC is the number of clusters. it is minus two because $d(i,i) = 0$ and $d(i,j)$ are not added in the summation in above formula so it subtract 2 to make it average of distance with the rest of clusters. Isolation function is not symmetric. $\gamma(i,j)$ is not equal with $\gamma(j,i)$. But the relative distance is symmetric and it considers the minimum of $\gamma(i,j)$ and $\gamma(j,i)$ in the denominator.

$$D_{rh}(i,j) = \frac{d(i,j)}{\min\{\gamma(i,j), \gamma(j,i)\}}$$

The disadvantage of ratio-type methods is that there can be branch crossing in the dendrogram in this type of method. It means that a level can have higher relative distance than the next level. While in distance type methods there is no branch crossing in the dendrogram.

3.3. Goodness of fit measure:

Two goodness-of-fit measures in this analysis was considered. These measures rely on the ranking of values when comparing the distances between two objects and their ultrametric distances within the hierarchical clustering. For a set of four values (a quadruple) to be considered as "concordant," it means that there is a consistent pattern in the signs when comparing the distances between them and their ultrametric distances. In other words, if we're comparing the distances of objects i and j to those of objects k and l, for them to be considered concordant:

- If the signs of both the distance comparisons (i to j and k to l) are the same, and

$$d(i,j) < d(k,l) \Rightarrow U(i,j) < U(k,l)$$

- If the signs of both the ultrametric distance comparisons (i to j and k to l) are also the same,

$$d(i,j) > d(k,l) \Rightarrow U(i,j) > U(k,l)$$

Then, these four values are considered concordant. Here d is Euclidean distance and U is ultrametric distance.

Consequently, the goodness-of-fit measures used are:

a. **Kendall's tau:** this measure considers all possible pairs of quadruples in calculating the goodness of fit for structure of the dendrogram. It is computed using

$$\tau = \frac{S^+ - S^-}{\frac{(n+1)n(n-1)(n-2)}{8}}.$$

b. **Goodman-Kruskal**: this measure for goodness of fit the number of non-comparable and tied observations are not considered in the denominator. It is computed using

$$GK = \frac{S^+ - S^-}{S^+ + S^-}$$

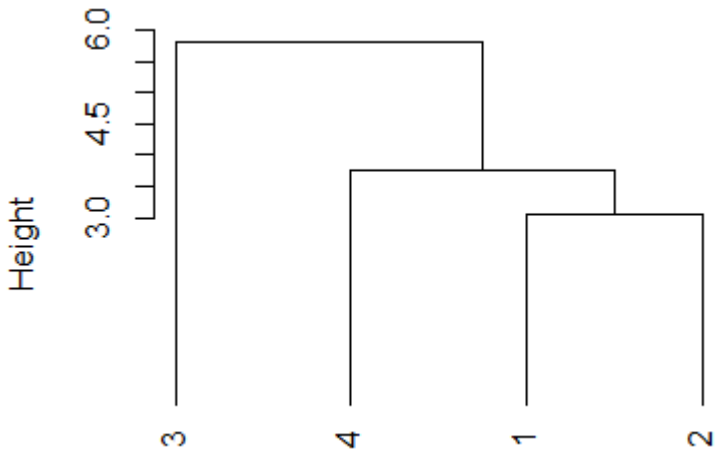

Figure 1: Dendrogram illustration for hierarchical clustering using four cases

Table 1: comparison of quadruples.

quadruple	i	J	k	l	d(i,j)	d(k,l)	u(i,j)	u(k,l)	status	S+	S-	nc	GK	Tau
1	1	2	1	3	3.04	5.80	3.04	5.80	concordant	10	1	4	0.8182	0.6667
2	1	2	1	4	3.04	3.35	3.04	3.76	concordant					
3	1	2	2	3	3.04	3.56	3.04	5.80	concordant					
4	1	2	2	4	3.04	3.76	3.04	3.76	concordant					
5	1	2	3	4	3.04	4.79	3.04	5.80	concordant					
6	1	3	1	4	5.80	3.35	5.80	3.76	concordant					
7	1	3	2	3	5.80	3.56	5.80	5.80	nc					
8	1	3	2	4	5.80	3.76	5.80	3.76	concordant					
9	1	3	3	4	5.80	4.79	5.80	5.80	nc					
10	1	4	2	3	3.35	3.56	3.76	5.80	concordant					
11	1	4	2	4	3.35	3.76	3.76	3.76	nc					
12	1	4	3	4	3.35	4.79	3.76	5.80	concordant					
13	2	3	2	4	3.56	3.76	5.80	3.76	discordant					
14	2	3	3	4	3.56	4.79	5.80	5.80	nc					
15	2	4	3	4	3.76	4.79	3.76	5.80	concordant					

In Figure 1, Dendrogram illustration for hierarchical clustering using four cases for example for four cases of x_1, x_2, x_3, x_4 there are in total (n+1)*(n)*(n-1)*(n-2)/8 = 5*4*3*2 / 8 = 15 cases for comparison. In the illustration dendrogram in table 1, the quadruple of objects (1,3) and (2,3) cannot be compared directly. This is because both cases 1 and 2 belong to the same cluster, and they are connected to object 3 at the same hierarchical level. Consequently, their ultrametric distances are the same. However, among these four cases, only one discordant quadruple can be identified, namely, (2,3) and (2,4). In this case, the discordance arises because the distance between object 2 and 3 is smaller than the distance between object 2 and 4. However, the ultrametric distance between object 2 and 4 is smaller than that between object 2 and 3. Kendall's Tau, a correlation coefficient, considers all 15 possible quadruples in its denominator when calculating its value. This means it considers all potential combinations for comparing the ranks. On the other hand, the Goodman-Kruskal index does not include non-comparable (cases where objects are at the same hierarchical level) and tied quadruples in its calculations. It focuses on the comparable cases, where meaningful comparisons can be made, and excludes those cases where ranking the objects is not feasible.

4. RESULTS

Hierarchical clustering performed on the Pottery dataset using various distance type methods, including average linkage, complete linkage, single linkage, median linkage, and the centroid method. Also the Mollineda & Vidal [9] relative hierarchical clustering as a ratio-type method was utilized. In total, there are 45 cases in this dataset, resulting in 489,555 possible quadruples for comparison. These clustering methods was applied, in conjunction with various scaling techniques. The Goodman-Kruskal measure of goodness of fit was calculated for each method combined with each scaling approach for this dataset, and the results can be found in Table 2. In a previous study by Roux [6], the mean absolute deviation was used for scaling this dataset. It was observed that Mollineda & Vidal [9] algorithm for relative hierarchical clustering yielded the highest Goodman-Kruskal measure (0.8066), indicating superior performance compared to other methods. However, the data in Table 2 shows that when the median linkage method used, a Goodman-Kruskal measure of 0.8072, surpassing the ratio-type method. Notably, across four out of five scaling methods used for this dataset, the median linkage method consistently outperforms other clustering techniques. Regarding the structure of the dendrogram, it is worth mentioning that when using median absolute deviation for scaling, the highest Goodman-

Kruskal value of 0.9233 is achieved with the average linkage method. Additionally, from the 489,555 quadruples examined, the highest number of concordant quadruples is observed with the single linkage method, while the lowest number of discordant quadruples is associated with the average linkage method.

Furthermore, when considering the Kendall's Tau measure, the single linkage method with a Tau value of 0.7191, using median absolute deviation for scaling, achieves the highest Tau measure. Following closely is the median linkage method with a Tau value of 0.7180. In terms of the average Goodman-Kruskal measure, the average linkage method performs best, and Mollineda & Vidal [9] method in combination with median absolute deviation scaling is the next best option. For other scaling techniques, the median linkage method consistently achieves the highest Goodman-Kruskal measure.

Table 2. Hierarchical clustering using Pottery dataset.

Scale	Average linkage			Complete linkage			Single linkage		
	S+	S-	G-K	S+	S-	G-K	S+	S-	G-K
mean absolute deviation	343726	37095	0.8052	334917	41646	0.7788	343999	38322	0.7995
median absolute deviation	366071	***14604***	***0.9233***	359224	18135	0.9039	***367244***	15223	0.9204
Standard deviation	340559	39958	0.7900	340587	40150	0.7891	341490	40759	0.7867
interquartile range	361508	19239	0.8989	361654	18911	0.9006	362943	19374	0.8986
None	338106	41195	0.7828	295928	54953	0.6868	338987	42652	0.7765
Scale	Mollineda & Vidal (2000)			Median linkage			Centroid		
	S+	S-	G-K	S+	S-	G-K	S+	S-	G-K
mean absolute deviation	343225	36750	0.8066	343765	36670	***0.8072***	344401	37344	0.8044
median absolute deviation	366070	14605	0.9233	366292	14815	0.9223	366202	15075	0.9209
Standard deviation	340582	39883	0.7903	340809	39626	***0.7917***	341439	39956	0.7905
interquartile range	361510	18427	0.9030	362536	17965	***0.9056***	362817	18792	0.9015
None	341701	41962	0.7813	338106	41195	***0.7828***	338132	41283	0.7824

Also, in the results presented in Table 2, the various methods evaluated by considering the diameter of the objects being connected. The diameter of objects in C_1 connected with objects in C_2 is defined as the maximum distance between any object in C_1 and any object in C_2. While the node level could be taken as the level of the dendrogram in each hierarchical clustering model, for comparison with ratio-type methods (which may involve branch crossing) and for consistency with the ultrametric distance calculation method used by Roux [6], diameter was used as the basis

for evaluation in other linkage methods as well. In addition, when the diameter for evaluating the hierarchical clustering models employed, it became evident that the method proposed by Mollineda & Vidal [9] did not perform better than the Median linkage method. However, using the node level itself for evaluation instead of the diameter, the distance-based methods demonstrate superior performance when compared to ratio-type methods.

In Table 3, the results of the goodness-of-fit assessment are presented using the node level of each model. For instance, the node level for the objects in C_1 connected with objects in C_2 in the case of average linkage is defined as the average of the distances between objects in C_1 and objects in C_2. Thus, in this context, the node level is used as an alternative to the diameter for evaluation. As shown in Table 3, when evaluating the clusters using the node level, the Goodman-Kruskal (GK) value for median linkage is the highest at 0.9287, particularly when median absolute deviation is employed for data scaling. The highest concordant quadruple is no longer observed for single linkage. Instead, the method with the most concordant quadruples is median linkage, where median absolute deviation equal 367514. This suggests that median linkage outperforms other linkage methods for this dataset. Comparing the results in Table 3 to those in Table 2, when mean absolute deviation is used for data scaling, average linkage with GK = 0.8082 performs better than Mollineda & Vidal [9] with GK = 0.8066. Additionally, Kendall's tau measure also indicates that median linkage (tau = 0.7229) exhibits the best performance, followed by average linkage (tau = 0.7221).

Table 3. Hierarchical clustering using Pottery dataset.

	Average linkage			Single linkage			Median linkage		
Scale	S+	S-	G-K	S+	S-	G-K	S+	S-	G-K
mean absolute deviation	344300	36521	0.8082	343921	38400	0.7991	344172	36263	***0.8094***
median absolute deviation	367092	***13583***	0.9286	367183	15284	0.9201	***367514***	13593	***0.9287***
Standard deviation	341125	39392	0.7930	340770	41479	0.7830	341391	39044	***0.7947***
interquartile range	363091	17656	0.9073	363297	19020	0.9005	363536	16965	***0.9108***
None	338648	40653	***0.7856***	338345	43294	0.7731	338642	40659	0.7856

Hierarchical clustering was conducted on the Fisher Irish dataset, which comprises 150 observations. Consequently, there are a total of 62,434,725 quadruples available for comparison within the dendrogram of the Fisher Irish dataset. With larger sample sizes, the number of quadruples significantly increases, making the calculation of the goodness-of-fit measures more time-consuming. In this dataset, around 62 million quadruples need to be compared to calculate

Kendall's Tau and the Goodman-Kruskal measure. As displayed in Table 4, the highest Goodman-Kruskal index (GK) is observed in the dataset without any scaling, where GK = 0.8696 for the Centroid method, followed by GK = 0.8669 for Median linkage. Notably, when scaling the data using the Median Absolute Deviation method, the results from Mollineda & Vidal [9] outperform other methods, achieving a GK of 0.7984. The quadruples with the highest concordance are found in the case of single linkage, with a pattern of 4441142. The lowest number of discordant quadruples is observed in the Centroid method, with 3,042,850 discordant quadruples. Table 4 also evaluates the methods based on the diameter of the objects being connected. The highest Kendall's Tau measure is achieved by Median linkage, with tau = 0.6579, followed by the Centroid method, with tau = 0.6500. When focus shifted from diameter to node level for evaluating hierarchical clustering methods, the Average linkage method emerges as the top performer with a GK of 0.8714, followed by the Centroid method with a GK of 0.8707. Median linkage shows the highest number of concordant quadruples at 44 million, while the lowest number of discordant quadruples is observed with the Average linkage method, totaling 2.99 million discordant quadruples.

Table 4. Hierarchical clustering using Fisher Irish dataset.

Scale	Average linkage			Complete linkage			Single linkage		
	S+	S-	G-K	S+	S-	G-K	S+	S-	G-K
mean abs dev	42086950	5225373	0.7791	37208588	7520098	0.6637	43471181	5808487	0.7643
median abs dev	42451060	5012538	0.7888	41934821	5523014	0.7672	43083838	6399420	0.7414
STD	42813393	4543039	***0.8081***	38315238	6616353	0.7055	43596639	5674558	0.7697
IQR	41695840	5825778	0.7548	40283546	8741015	0.6434	42242358	7336471	0.7040
None	43484474	3121120	0.8661	39266049	5525015	0.7533	***44441142***	4040738	0.8333
Scale	Mollineda & Vidal (2000)			Median linkage			Centroid		
	S+	S-	G-K	S+	S-	G-K	S+	S-	G-K
mean abs dev	42685550	4834241	0.7965	42090010	4580544	***0.8037***	42848616	4698038	0.8024
median abs dev	42397282	4751468	**0.7984**	42007397	5186843	0.7802	42862557	5177243	0.7845
STD	42756263	4773238	0.7991	42895673	4691526	0.8028	42979095	4719100	0.8021
IQR	41403483	5787590	0.7547	41830925	5718549	***0.7595***	42229718	6329571	0.7393
None	43317864	3205491	0.8622	44232999	3154613	0.8669	43627528	***3042850***	***0.8696***

The Goodman-Kruskal (GK) value for Mollineda & Vidal [9] is also lower when compared to Table 5. This difference in the GK value is primarily due to the presence of branch crossings in ratio-type methods. When using the node level itself to evaluate the method, it can result in a

reduction in the goodness of fit measure. Among the goodness-of-fit measures, Kendall's tau demonstrates the best performance. Specifically, it shows the highest performance in Median linkage with a tau value of 0.6597. Following that, the Centroid method performs well with a tau value of 0.6508.

Table 5. Hierarchical clustering using Fisher Irish dataset.

Scale	Average linkage			Complete linkage			Single linkage		
	S+	S-	G-K	S+	S-	G-K	S+	S-	G-K
None	43609353	***2996474***	***0.8714***	39266049	5525015	0.7533	43825184	4610803	0.8096
Scale	Mollineda & Vidal (2000)			Median linkage			Centroid		
	S+	S-	G-K	S+	S-	G-K	S+	S-	G-K
None	43110739	3419164	0.8530	***44289757***	3098817	0.8692	43652547	3017939	0.8707

5. DISCUSSION

The focus of this paper was on hierarchical clustering, which involved the use of various linkage methods to compare the outcomes of distance-based methods with a ratio-based method introduced by Mollineda & Vidal [9]. This ratio-based method was identified as the most effective approach in a previous study conducted by Roux [6]. To facilitate this comparison, different scaling techniques employed on the same real datasets that were used in Roux's earlier study. The findings revealed that the results obtained from the distance-based methods were not inferior to those from the ratio-based method. Contrary to the results reported by Roux [6], the performance of the distance-based methods outperformed the ratio-based method in both real datasets. This conclusion was supported by both the Goodman-Kruskal and Kendall's Tau measures of goodness of fit. For the Pottery dataset, the Median Absolute Deviation yielded the best results for scaling the dataset, while for the Fisher Irish dataset, the highest goodness-of-fit measures were obtained when no scaling was applied. Specifically, in the Pottery dataset, the Median linkage method performed best, followed by the Average linkage method. In the case of the Fisher Irish dataset, the Median linkage method also showed the highest performance according to Kendall's Tau, followed by the centroid method. In the evaluation of the clustering structures using diameter as a criterion, the centroid method proved to be the most effective with a Goodman-Kruskal value of 0.8696, followed closely by the Median linkage method with a Goodman-Kruskal value of 0.8669. On the other hand, when the structures using node level was assessed, the Average linkage method delivered the best performance with a Goodman-Kruskal

value of 0.8714, with the centroid method as the runner-up with a Goodman-Kruskal value of 0.8707.

6. CONCLUSION

In summary, the study demonstrated that the distance-based methods outperformed the ratio-based method proposed by Mollineda & Vidal [9] in both datasets. Additionally, it has been found that the choice of data scaling method significantly influenced the hierarchical clustering results and selecting the appropriate scaling method for each dataset led to more consistent clustering structures.

Conflicts of Interest: The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] A.K. Jain, Data Clustering: 50 Years Beyond K-Means, Pattern Recognit. Lett. 31 (2010), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011.

[2] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, A. Song, Efficient Agglomerative Hierarchical Clustering, Expert Syst. Appl. 42 (2015), 2785-2797. https://doi.org/10.1016/j.eswa.2014.09.054.

[3] H. Mittal, A.K. Tripathi, A.C. Pandey, P. Venu, V.G. Menon, R. Pal, A Novel Fuzzy Clustering-Based Method for Human Activity Recognition in Cloud-Based Industrial IoT Environment, Wireless Netw. (2022). https://doi.org/10.1007/s11276-022-03011-y.

[4] M.G. Kendall, A New Measure of Rank Correlation, Biometrika. 30 (1938), 81-93. https://doi.org/10.2307/2332226.

[5] L. Goodman, W. Kruskal, Measures of Association for Cross-Validations, Part 1, J. Amer. Stat. Assoc. 49 (1954), 732–764.

[6] M. Roux, A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms, J. Classif. 35 (2018), 345-366. https://doi.org/10.1007/s00357-018-9259-9.

[7] A. Tubb, A.J. Parker, G. Nickless, The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry, Archaeometry. 22 (1980), 153-171. https://doi.org/10.1111/j.1475-4754.1980.tb00939.x.

[8] R.A. Fisher, The Use of Multiple Measurements in Taxonomic Problems, Ann. Eugenics. 7 (1936), 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[9] R.A. Mollineda, E. Vidal, A Relative Approach to Hierarchical Clustering. In: Pattern Recognition and Applications, vol. 56, pp. 19–28. IOS Press, Amsterdam (2000).