Comparison of the Efficiency of Linkage in Hierarchical Cluster Analysis for Multivariate Data

Kullanat Phongduang [a], Nutchanon Buasuwan [b], Pakhamon Kittikunsuntorn[c], Pattarapong Chanaarpa [d]

[a]Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

Abstract

Despite the wide use of hierarchical clustering, the selection of an appropriate linkage method remains a crucial factor that directly affects clustering performance. In this study, several linkage methods, namely Average, Centroid, Complete, McQuitty, Median, Single, and Ward, were compared for clustering multivariate datasets. In this research, the analysis was conducted using both simulated and real-world datasets. The simulated datasets were generated to represent various data structures. Additionally, a real dataset was used to examine further whether the results from the simulated data were consistent with real-world situations. To compare the clustering methods, several evaluation metrics were used, including accuracy, precision, recall, and F-score. The results show that the Ward method provided the best clustering performance across all datasets used in the experiments. Similarly, the results from the real-world dataset also indicated that the Ward method achieved the best clustering performance. These findings suggest that Ward provides the most effective and reliable clustering performance for hierarchical clustering in multivariate datasets.

Keywords: Hierarchical clustering, Agglomerative, Linkage Methods, Ward Method, Performance metrics

1. Introduction

Clustering is an unsupervised machine learning technique used to group data points or observations into clusters based on their similarities or shared patterns [1]. It is widely applied in exploratory data analysis to uncover hidden structures within data. In general, clustering methods can be broadly classified into two main types: Hierarchical Clustering Analysis (HCA) and K-means clustering. Among these, hierarchical clustering is a commonly used approach that organizes data into a tree-like structure, allowing relationships between observations to be interpreted at different levels [2].

Hierarchical clustering can be divided into two main approaches: agglomerative and divisive methods. Agglomerative hierarchical clustering is one of the most widely used techniques, following a bottom-up strategy in which each observation initially forms its own cluster. These clusters are then gradually merged based on their similarity until a single cluster is formed [3]. A key component of this method is the linkage criterion, which determines how the distance between clusters is calculated. As different linkage methods define cluster similarity in different ways, they can produce noticeably different clustering results. In addition to linkage methods, the choice of distance metric plays an important role in shaping the clustering outcome. In this study, Euclidean distance is employed due to its simplicity and effectiveness in measuring similarity between observations [4]. However, the results of hierarchical clustering can vary significantly depending on both the distance metric and the linkage method used [5]. This makes the selection of an appropriate linkage method an important and non-trivial task, especially when the underlying structure of the data is not clearly known. Several linkage methods have been proposed, each with its own strengths and limitations. Single linkage defines the distance between clusters based on the minimum distance between observations, which makes it effective in detecting chain-like patterns but also prone to the chaining effect [6]. Complete linkage, in contrast, uses the maximum distance and tends to produce more compact clusters, although it can be sensitive to outliers [7]. Average linkage computes the average pairwise distance between clusters, providing a balance between these two approaches [8]. Other methods, such as centroid and median linkage, determine distances based on measures of central tendency. While these methods are intuitive, they may lead to reversals in the dendrogram, complicating interpretation [9]. McQuitty linkage applies a simple averaging scheme that does not account for cluster size, which may introduce bias in some cases [10]. Ward linkage takes a different approach by minimizing the increase in within-cluster variance during the merging process, often producing compact and internally consistent clusters [11].

Given these differences, the choice of linkage method can strongly influence the resulting cluster structure. Inappropriate selection may lead to misleading patterns and reduce the reliability of the analysis. Although many methods are available, there is still no clear consensus on which approach performs best across different types of data. Among the available methods, Ward linkage has been widely recognized for its effectiveness. By minimizing within-cluster variance at each step, it tends to produce clusters that are more compact, stable, and easier to interpret. Previous studies have shown that Ward linkage often provides more

balanced and meaningful clustering results compared to methods such as single or complete linkage [7], [8]. These advantages have made it one of the most commonly used linkage methods in hierarchical clustering.

To evaluate clustering performance, this study employs external validation measures, including accuracy, precision, recall, and F-score. These metrics are computed by comparing the clustering results with predefined class labels in the dataset. Although clustering is an unsupervised technique, such measures can be applied when ground truth labels are available, allowing for a more objective and quantitative comparison of clustering outcomes [12].

Therefore, this study aims to compare the performance of seven linkage methods in agglomerative hierarchical clustering and to identify the most appropriate methods for different data structures. The findings are expected to provide practical guidance for selecting suitable linkage methods in real-world applications.

The study starts with a description of the datasets and experimental design, followed by a detailed presentation of the methodology and theoretical framework. Then, the results are presented and discussed, and the paper concludes with a summary of the main findings.

2.Method

In this study, labeled datasets were used to evaluate the performance of hierarchical clustering. The data were divided into two types: simulated datasets and real-world datasets. Euclidean distance was used as the distance measure to calculate the similarity between data points. Seven linkage methods were applied, including Single, Complete, Average, Centroid, Median, Ward, and McQuitty linkage, in order to compare the performance of each method. The clustering performance was evaluated using standard performance metrics, namely accuracy, precision, recall, and F-score.

Table 1 Summary for datasets D1–D5 and Wine

Dataset	Number of Variables	Number of Groups	Total Observations	Number of Observations in Each Group
D1	3	4	1000	[244,255,235,266]
D2	5	4	996	[240,260,232,264]
D3	12	4	990	[255,242,251,242]
D4	15	3	988	[340,308,340]
D5	20	4	989	[244,281,236,228]
Wine	13	3	172	[59,66,47]

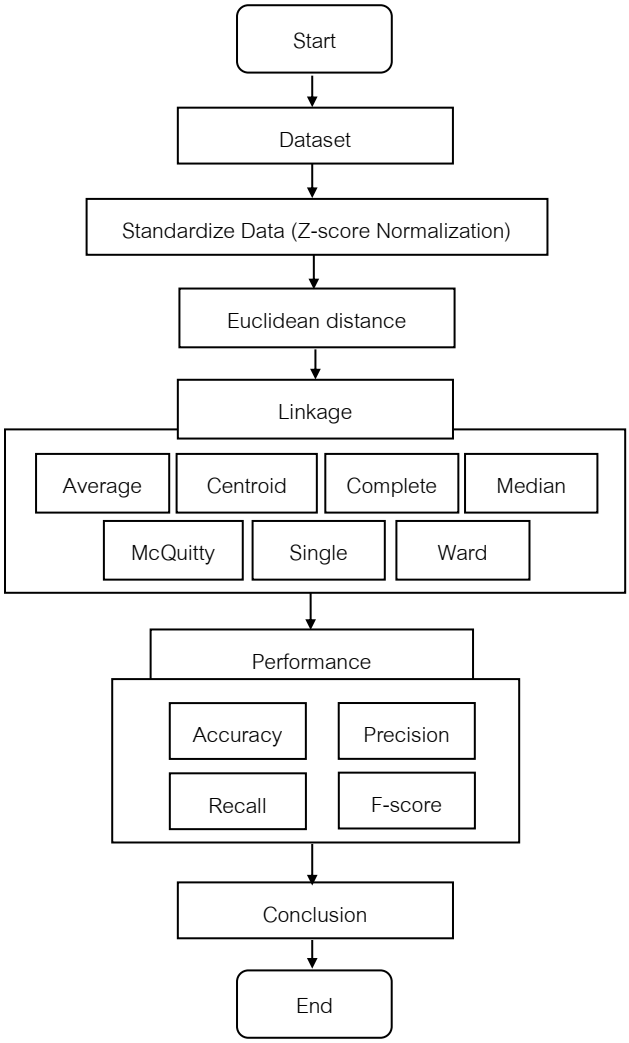


Fig 1. Flowchart of the Experimental Procedure

In this study, labeled datasets, including both simulated and real-world data, were used to evaluate the performance of agglomerative hierarchical clustering. All data were standardized using Z-score normalization before analysis. Seven linkage methods Single, Complete, Average, Centroid, Median, McQuitty, and Ward-were applied to each dataset. The clustering performance was evaluated using external validation measures, including accuracy, precision, recall, and F-score, by comparing the results with the true class labels. Finally, the performance of each method was compared to determine the most suitable linkage approach for different data structures.

3. Methodology

In this study, relevant theories, concepts, and previous research are presented to provide a conceptual framework for the research. The focus is on cluster analysis and hierarchical cluster analysis, which are techniques within unsupervised learning. These methods aim to discover underlying patterns and structures in data without requiring predefined group labels. This theoretical foundation supports the selection of appropriate analytical methods for the data used in the study.

Hierarchical Cluster Analysis (HCA) is an unsupervised machine learning technique used to explore and group similar data into a hierarchical structure [1]. A distinctive feature of this analytical method is its systematic and deterministic approach to clustering. Specifically, once a data point has been assigned to a

particular cluster, it cannot be reassigned to another. This irreversibility distinguishes HCA from non-hierarchical (partitional) clustering methods, which allow for the dynamic reassignment of data points to optimize the cluster formations [13].

3.1 Hierarchical Clustering

Hierarchical cluster analysis can be categorized into two primary methodological approaches

1. Agglomerative Method (Bottom-up) This technique begins by treating each individual data point as a separate, independent cluster. It then iteratively pairs and merges the most similar (or closest) clusters in a step-by-step hierarchy until all data points are eventually aggregated into a single overarching cluster [1].

2. Divisive Method (Top-down) Operating in the reverse direction, this technique starts with all data points encompassed within a single large cluster. It proceeds by successively partitioning the data into smaller sub-clusters based on their maximum dissimilarity, continuing this division until each data point forms its own distinct singleton cluster [1].

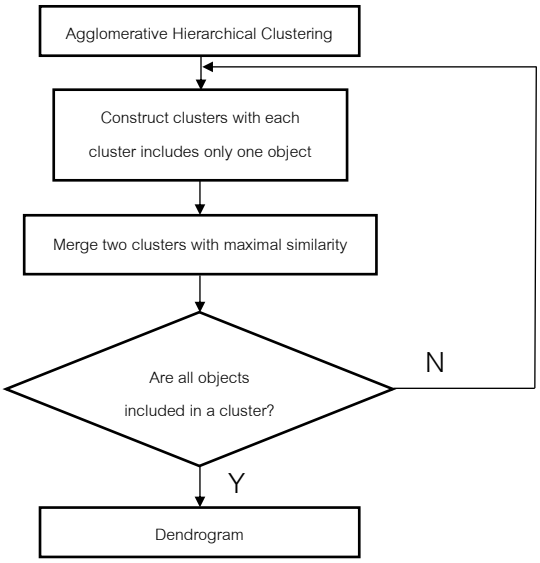


Fig 2. The flowchart of agglomerative hierarchical clustering

In this study, hierarchical cluster analysis was conducted utilizing the agglomerative technique. The agglomerative approach is a "bottom-up" clustering process that initially treats each individual data unit as an isolated cluster. Subsequently, it iteratively merges the most similar or closely related clusters until a hierarchically structured relationship among the clusters is ultimately formed

3.2 Distance measure

In this study, measuring the distance between data points is an important step in determining how similar or different the observations are before performing clustering. Among the available distance measures, Euclidean distance is one of the most commonly used because it is straightforward and easy to interpret. Many studies have adopted Euclidean distance as a standard metric in clustering analysis due to its effectiveness in

representing the actual distance between points in multidimensional space [2]. Therefore, Euclidean distance was used in this study.

Let D_{ij} represent the distance between observation I and observation j, where $i \neq j$ The distance is calculated as follows

$$D_{ij} = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk} \right)^2} \quad \text{....(1.1)}$$

3.3 Methods for Combining Clusters

In hierarchical clustering analysis, the final grouping results are significantly influenced by the selection of an appropriate linkage method. This research evaluates and compares seven distinct linkage criteria as follows

1. Single Linkage

This method determines the distance between two clusters based on the minimum distance between the nearest pair of members from each cluster[1].

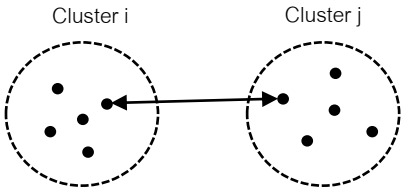


Fig 3. Single Linkage

2. Complete Linkage

This method defines the distance between clusters based on the maximum distance between the furthest pair of members from each cluster[1].

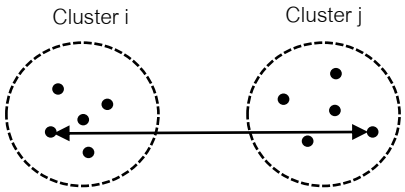


Fig 4. Complete Linkage

3. Average Linkage

This method calculates the average distance between all possible pairs of observations between the two clusters[1].

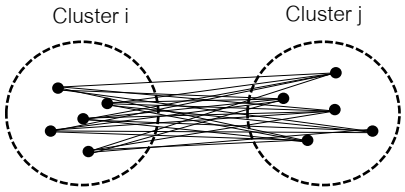


Fig 5. Average Linkage

4. Centroid Linkage

This method measures the distance between the geometric centers (centroids) of each cluster. When two clusters are merged, the new centroid is calculated as the average of all data points within the newly formed group[1].

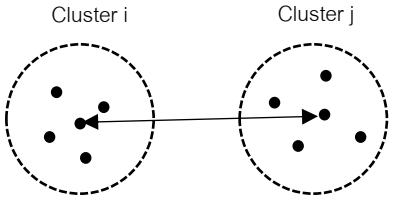


Fig 6. Centroid Linkage

5. Median Linkage

This method calculates the distance from the median point of the merged clusters. It utilizes a weighting scheme that assigns equal importance to each cluster regardless of its size [1].

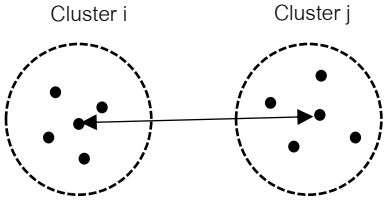


Fig 7. Median Linkage

6. McQuitty Linkage

This method employs a weighted average calculation to determine the relationship between merging clusters, assuming that the distance to an external cluster is the average of the distances from its constituent sub-clusters [1].

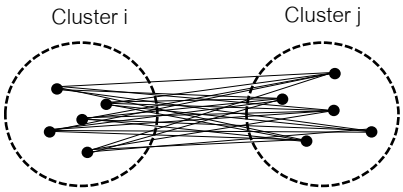


Fig 8. McQuitty Linkage

7. Ward Linkage

This method focuses on minimizing the increase in total within-cluster variance at each step. It merges clusters that result in the smallest possible increase in the Error Sum of Squares (ESS) [1].

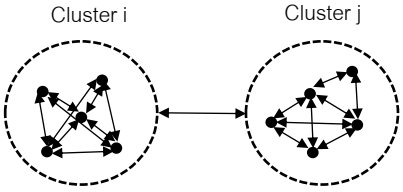


Fig 9. Ward Linkage

3.4 Dendrogram

A dendrogram is used to illustrate how clusters are formed, where the horizontal lines represent the distance between clusters. Longer lines indicate greater differences and can help determine an appropriate stopping point. When lines appear close together across stages, it suggests a relatively stable level of homogeneity. Therefore, the cut-off point should be selected where lines are not densely grouped, while avoiding large increases in distance, to obtain a suitable number of clusters [14].

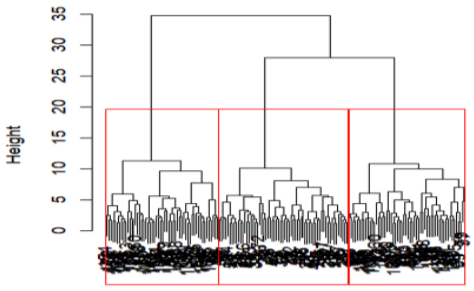


Fig 10. Dendrogram of Ward Linkage

3.5 Confusion Matrix

A confusion matrix is a table used to summarize the predictive performance of a classification model by comparing the model's predicted values against the actual values found in the dataset. As a widely adopted standard tool for evaluating classification efficiency, this matrix provides a clear visualization of how the model performs, specifically highlighting where it makes correct predictions and where various types of errors occur [15].

Table 2 Confusion Matrix

		Actual Value	
		Positive	Negative
Predicted Values	Positive	True positive	False positive
	Negative	False negative	True negative

- True Positive (TP) is when the prediction is correct: the model predicts positive and the actual outcome is positive.
- True Negative (TN) is when the prediction is correct: the model predicts negative and the actual outcome is negative.
- False Positive (FP) is when the prediction is incorrect: the model predicts positive, but the actual outcome is negative.
- False Negative (FN) is when the prediction is incorrect: the model predicts negative, but the actual outcome is positive.

1. Accuracy

Accuracy is often used to measure how well a classification model performs because it tells us how many predictions were correct out of all the observations in the dataset [15]. To calculate accuracy, you divide the number of correct predictions by the total number of samples. The equation below explains this.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

2. Precision and Recall

Precision and recall are useful metrics for telling apart different kinds of errors in data classification. They help us judge how well a classification model performs, not just by looking at accuracy, which only shows the overall rate of correct results. Accuracy gives a general overview, but precision and recall provide more detail by showing how the model handles positive cases and what happens when it makes mistakes. [15]

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

3. F-score

The F-score is useful because it brings together Precision and Recall. Using both gives a better sense of how well a model performs overall. Precision tells you how accurate the positive predictions are, and Recall shows how well the model finds all real positive cases. The F-score is especially helpful when you want to balance these two factors [15].

$$\text{F-score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision}+\text{Recall})}$$

4. Result and Discussion

The analysis results are presented in two main parts: (1) visualization of the clustering results and the characteristics of clusters obtained from each linkage method, and (2) comparison of the performance of each method using evaluation metrics, including accuracy, precision, recall, and F-score, in order to determine which method provides the highest clustering performance.

In addition, to enhance the reliability of the analysis, real-world data were used to support the findings obtained from synthetic data. The results from both types of datasets were tested and compared to provide more comprehensive and empirical conclusions regarding the performance trends of each linkage

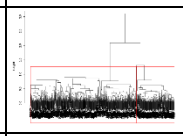
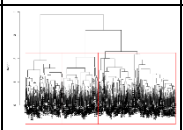
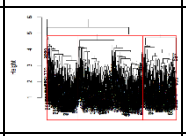
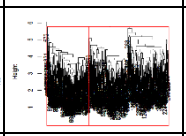
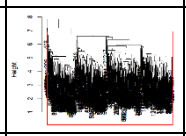
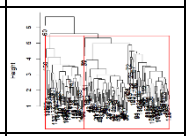
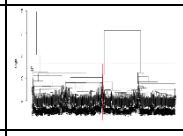
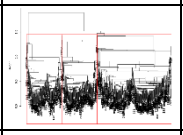
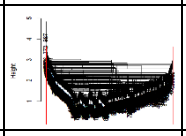
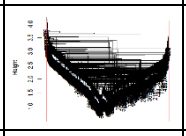
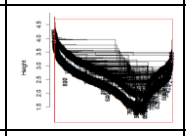
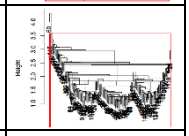
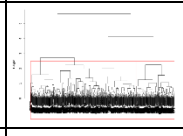
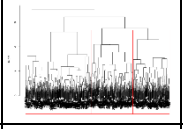
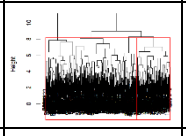
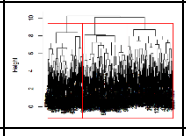
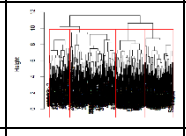
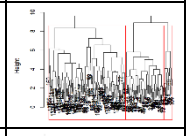
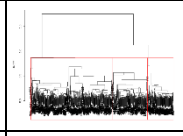
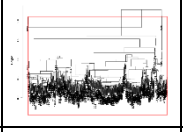
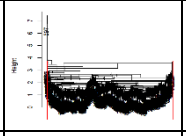
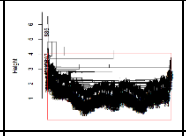
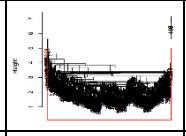
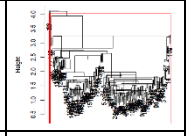
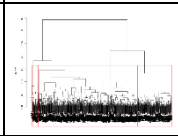
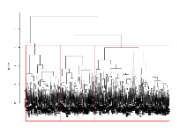
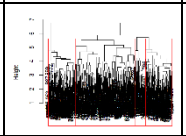
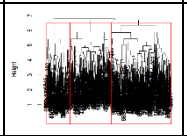
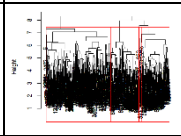
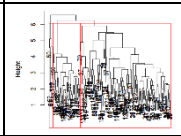
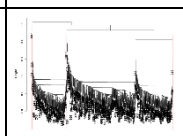
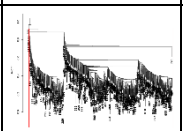
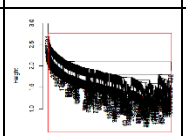
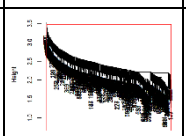
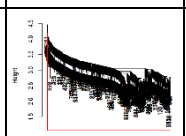
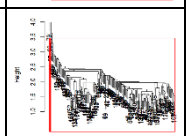
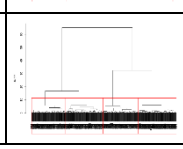
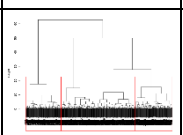
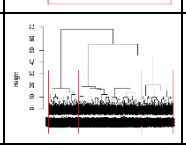
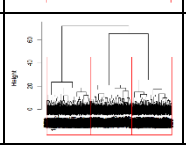
Dataset	D1	D2	D3	D4	D5	Wine
Average						
Centroid						
Complete						
Median						
McQuitty						
Single						
Ward						

Fig 11. Dendrograms of hierarchical clustering methods for datasets D1–D5 and Wine

The hierarchical clustering results for both the simulated and real (Wine) datasets show that the dendrogram patterns are similar across different linkage methods. Ward linkage stands out by clearly separating clusters and keeping the groups balanced with little overlap between them.

In contrast, the Single linkage method exhibits a "chaining effect", where data points are merged sequentially into a few disproportionately large clusters, resulting in poorly defined boundaries.

When comparing the simulated data to the real data, the main trends are similar. Ward and Complete linkage both give more stable and clear clusters than the other methods. In contrast, Centroid and Median methods often create uneven clusters and unpredictable merging patterns.

Table 3 Comparison of correct classification results across clustering linkage methods with reference values

Linkage	Dataset					
	D1 [244,255,235,266]	D2 [242,253,237,253]	D3 [255,242,251,242]	D4 [340,308,340]	D5 [244,281,236,228]	Wine [59,66,47]
Average	[244,255,0,263]	[239,244,231,253]	[252,2,249,235]	[338,1,337]	[1,277,0,0]	[59,1,47]
Centroid	[244,255,1,266]	[239,252,0,253]	[255,1,0,1]	[1,1,340]	[1,279,0,0]	[0,64,0]
Complete	[244,180,235,266]	[240,113,192,253]	[253,118,249,237]	[333,274,335]	[148,271,225,227]	[51,12,47]
Median	[244,253,0,181]	[1,1,125,253]	[255,0,1,1]	[339,0,1]	[1,280,1,0]	[0,64,0]
McQuitty	[244,255,43,263]	[242,26,89,253]	[53,212,250,215]	[317,182,337]	[243,19,220,226]	[59,5,37]
Single	[244,255,1,266]	[0,1,1,253]	[255,0,0,1]	[1,1,340]	[0,279,1,0]	[0,64,0]
Ward	[244,250,229,263]	[240,207,234,253]	[248,229,250,238]	[322,302,338]	[233,259,222,226]	[59,56,47]

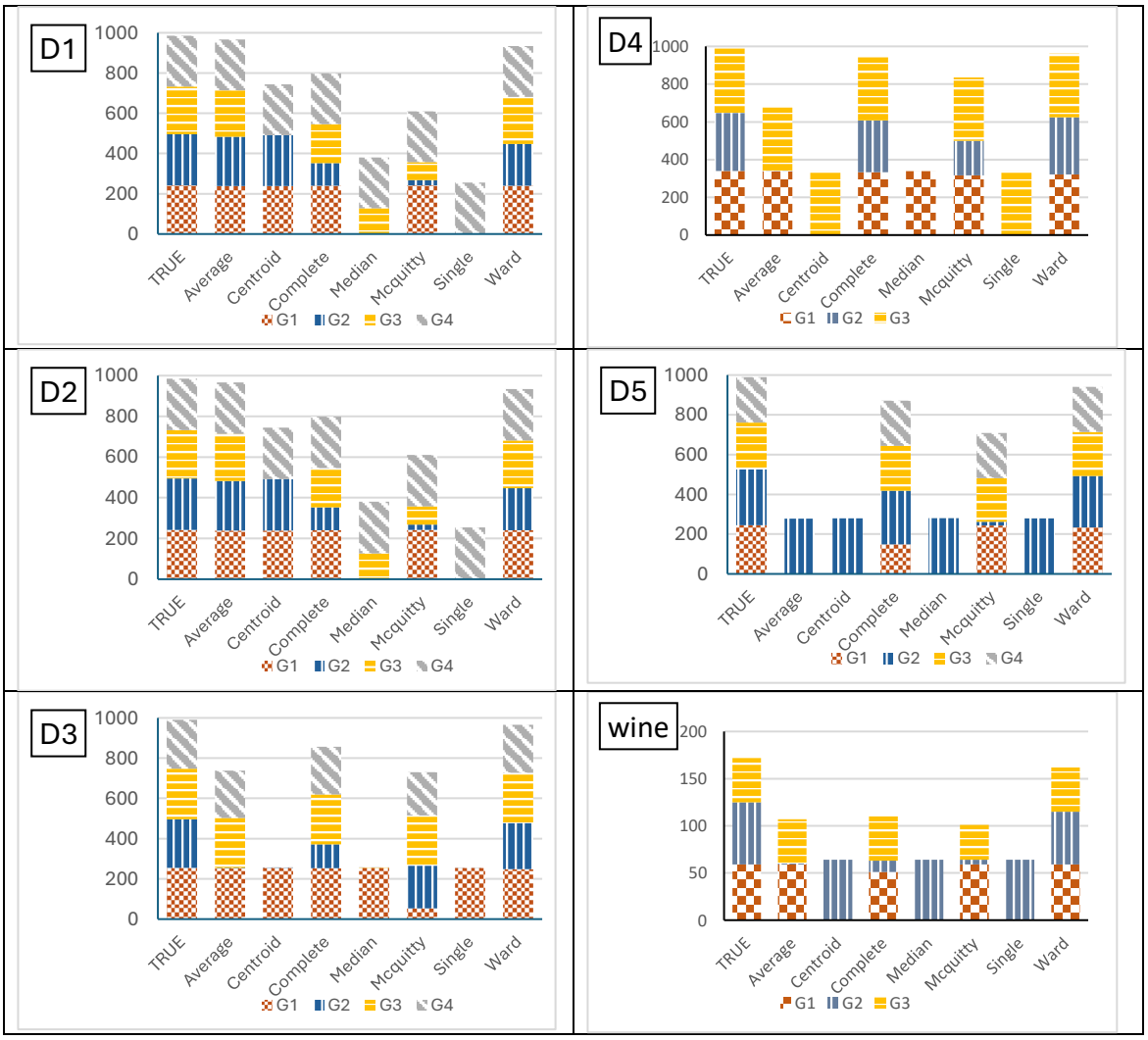


Fig 12. Bar chart comparing correct classification results across clustering linkage methods with reference values

Table 4 Overall performance comparison of linkage methods for simulated D1-D5 and Wine

Dataset	Performance	Linkage						
		Average	Centroid	Complete	Median	Mcquitty	Single	Ward
D1	Accuracy	0.7620	0.7660	0.9250	0.6780	0.8050	0.7660	0.9860
	Recall	0.6301	0.6304	0.7500	0.6104	0.6417	0.6304	0.7413
	Precision	0.7472	0.7511	0.9265	0.6682	0.7929	0.7511	0.9859
	F-score	0.6837	0.6854	0.8289	0.6380	0.7093	0.6854	0.8463
D2	Accuracy	0.9678	0.7309	0.8022	0.2962	0.7329	0.2661	0.9859
	Recall	0.7299	0.3737	0.6023	0.3182	0.3762	0.3162	0.7471
	Precision	0.9697	0.7460	0.8041	0.2875	0.7479	0.2511	0.9867
	F-score	0.8329	0.4979	0.6888	0.3020	0.5006	0.2799	0.8503
D3	Accuracy	0.7455	0.2596	0.8657	0.2596	0.7374	0.2586	0.9747
	Recall	0.8675	0.5646	0.9061	0.5646	0.7924	0.3146	0.9753
	Precision	0.7399	0.2521	0.8628	0.2520	0.7421	0.2510	0.9746
	F-score	0.7987	0.3485	0.8839	0.3485	0.7664	0.2792	0.9749
D4	Accuracy	0.6842	0.3462	0.9534	0.3441	0.8462	0.3462	0.9737
	Recall	0.8372	0.7816	0.9566	0.4479	0.8836	0.7816	0.9735
	Precision	0.6628	0.3354	0.9514	0.3333	0.8381	0.3354	0.9739
	F-score	0.7399	0.4694	0.9540	0.3822	0.8603	0.4694	0.9737
D5	Accuracy	0.2811	0.2831	0.8807	0.2851	0.7159	0.2831	0.95055
	Recall	0.3204	0.3207	0.9028	0.5710	0.8501	0.3207	0.9539
	Precision	0.2477	0.2492	0.8710	0.2512	0.7467	0.2493	0.9521
	F-score	0.2792	0.2805	0.8913	0.3489	0.7951	0.2805	0.9530
Wine	Accuracy	0.6221	0.3721	0.6395	0.3721	0.5872	0.3721	0.9419
	Recall	0.6717	0.3232	0.6821	0.3232	0.6210	0.3232	0.9495
	Precision	0.7956	0.1255	0.8005	0.1255	0.7560	0.1255	0.9407
	F-score	0.7284	0.1808	0.7366	0.1808	0.6819	0.1808	0.9451

Based on the results shown in Table 4, Fig 13 presents a visual comparison of the performance of different linkage methods on the simulated and Wine datasets.

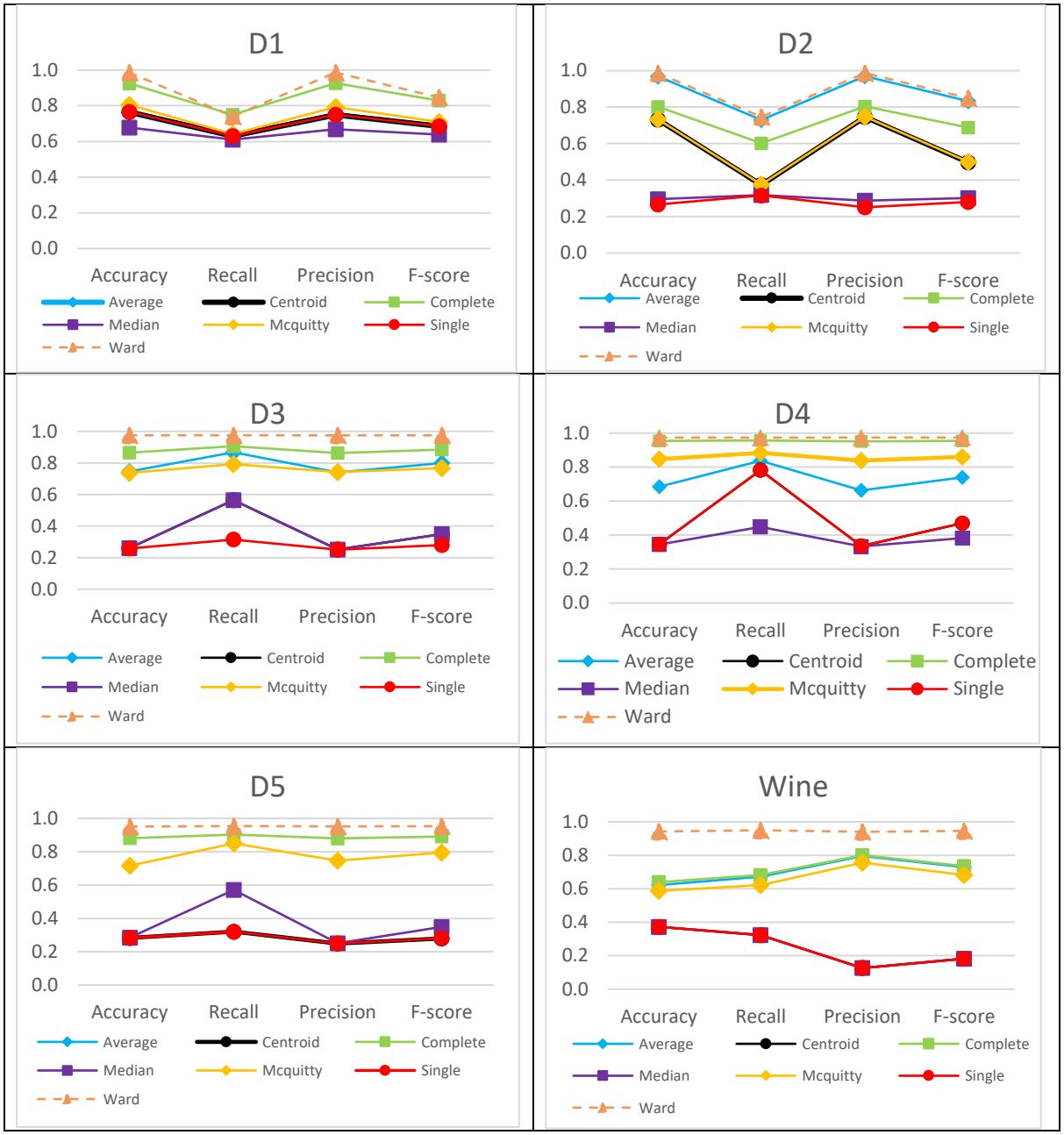


Fig 13. Performance comparison of linkage methods for dataset D1-D5 and Wine

Overall, the graphs show that Ward linkage works better than all the other methods for both datasets. Single linkage and Median linkage have the lowest performance. Complete and Average linkage perform at a moderate and steady level. These results suggest that Ward method is the best and most reliable method for this clustering task.

5.Conclusion

This study focuses on comparing several hierarchical clustering linkage methods using multiple datasets, including the Wine dataset and simulation datasets (D1–D5). The evaluation was carried out using Accuracy, Recall, Precision, and F-score to better understand how each method behaves under different data conditions.

From the results, Ward linkage stands out as the most consistent method. It tends to produce more compact clusters by controlling variation within each group, which helps improve overall clustering quality. This behavior can be observed across all datasets, suggesting that Ward is less affected by differences in data structure.

Other methods such as Complete, Average, and McQuitty also show acceptable performance, but their results are less stable. In some cases, their performance becomes comparable to Ward. For example, in dataset D2, the Average method gives very similar results, which may be related to the relatively even spread of the data. In dataset D4, the Complete method performs at a similar level, possibly because the clusters are clearly separated and tightly grouped. Even so, these patterns do not appear consistently across all datasets. In contrast, Centroid and Median methods tend to produce weaker results. Since both approaches rely on cluster centers, their performance can be affected when the data distribution is uneven or irregular, leading to less accurate grouping.

Single linkage shows the least reliable performance among all methods. While it may sometimes achieve a higher Recall, its Precision and F-score remain low. This is likely due to the chaining effect, where clusters are formed step by step by linking nearby points, often resulting in stretched and poorly defined clusters.Taking everything into account, Ward linkage provides the most balanced and dependable performance across different datasets. Other methods may still be useful in specific situations, but they are generally less consistent and are better suited for comparison rather than as a primary choice. Single linkage has the weakest performance overall. Sometimes it gets higher Recall, but its Precision and F-score are usually low. This shows problems like the chaining effect, where clusters are merged incorrectly and clustering quality suffers.

In summary, Ward linkage is the most effective and reliable method across different datasets. Other methods can be useful in some cases, but Ward should be the main choice. The other methods are best used for comparison, depending on the data.

References

1. IBM. (n.d.). *What is clustering?* IBM Think. https://www.ibm.com/think/topics/clustering
2. IBM. (n.d.). *What is hierarchical clustering?* IBM Think. https://www.ibm.com/think/topics/hierarchical-clustering
3. Brown, J. (2023). *Overview of agglomerative hierarchical clustering methods. British Journal of Computer, Networking and Information Technology.* https://abjournals.org/bjcnit/papers/volume-7/issue-2/overview-of-agglomerative-hierarchical-clustering-methods/
4. Iris Publishers. (n.d.). *Hierarchical clustering analysis (ABBA.MS.ID.000596).* https://irispublishers.com/abba/pdf/ABBA.MS.ID.000596.pdf
5. van Eck, N. J., & Waltman, L. (2011). *Citation-based clustering of publications.* arXiv. https://arxiv.org/pdf/1105.0121
6. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data.*
7. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
8. Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis.* Wiley.
9. Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson
10. Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38*, 1409–1438.
11. Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845
12. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press.
13. Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology, 11*(1), 8–21. https://doi.org/10.20982/tqmp.11.1.p008
14. Bratchell, N. (1989). Cluster analysis. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 1, pp. 55–64). Wiley.
15. Sathyanarayanan, S., & Tantri, S. (2024). *Confusion matrix-based performance evaluation metrics*. African Journal of Biomedical Research.