

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

Kullanat Phongduang, Nuchanon Buasuwan, Pakhamon Kittikunsuntorn, Pattarapong Chanaarpa, Chantha Wongoutong  
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.

## Introduction

Clustering is an unsupervised machine learning technique used to group data based on similarity and uncover hidden patterns. Hierarchical clustering is widely used because it represents data relationships in a tree-like structure. However, its results strongly depend on the choice of linkage method and distance metric, which can lead to significantly different outcomes. Although many linkage methods exist, there is still no clear consensus on which performs best across different datasets. Therefore, this study compares seven linkage methods using accuracy, precision, recall, and F-score to identify the most suitable methods and provide guidance for practical applications.

## Objective

- To study linkage methods in hierarchical clustering analysis, including Single, Complete, Average, Centroid, Median, Ward, and McQuitty methods, under the Euclidean distance measure.
- Compare the performance of the linkage methods using clustering performance evaluation metrics, namely accuracy, precision, recall, and F-score.

## Materials

This study evaluates hierarchical clustering using labeled datasets, including both simulated and real-world data. Euclidean distance is applied to measure similarity, and seven linkage methods (Single, Complete, Average, Centroid, Median, Ward, and McQuitty) are compared. The performance is assessed using 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 Observation in Each Group
D1	4	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]

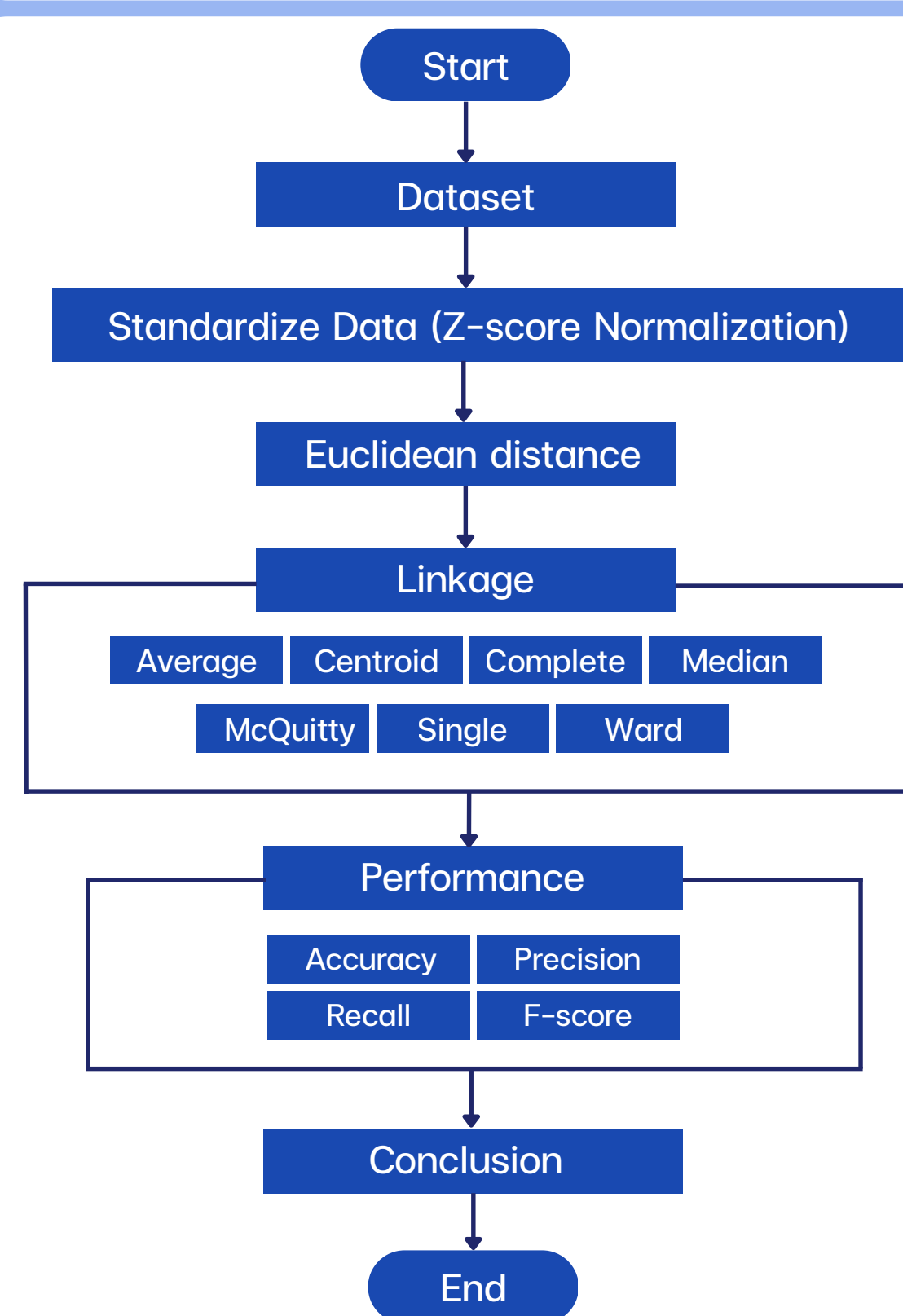


Fig 1. Flowchart of the Experimental Procedure

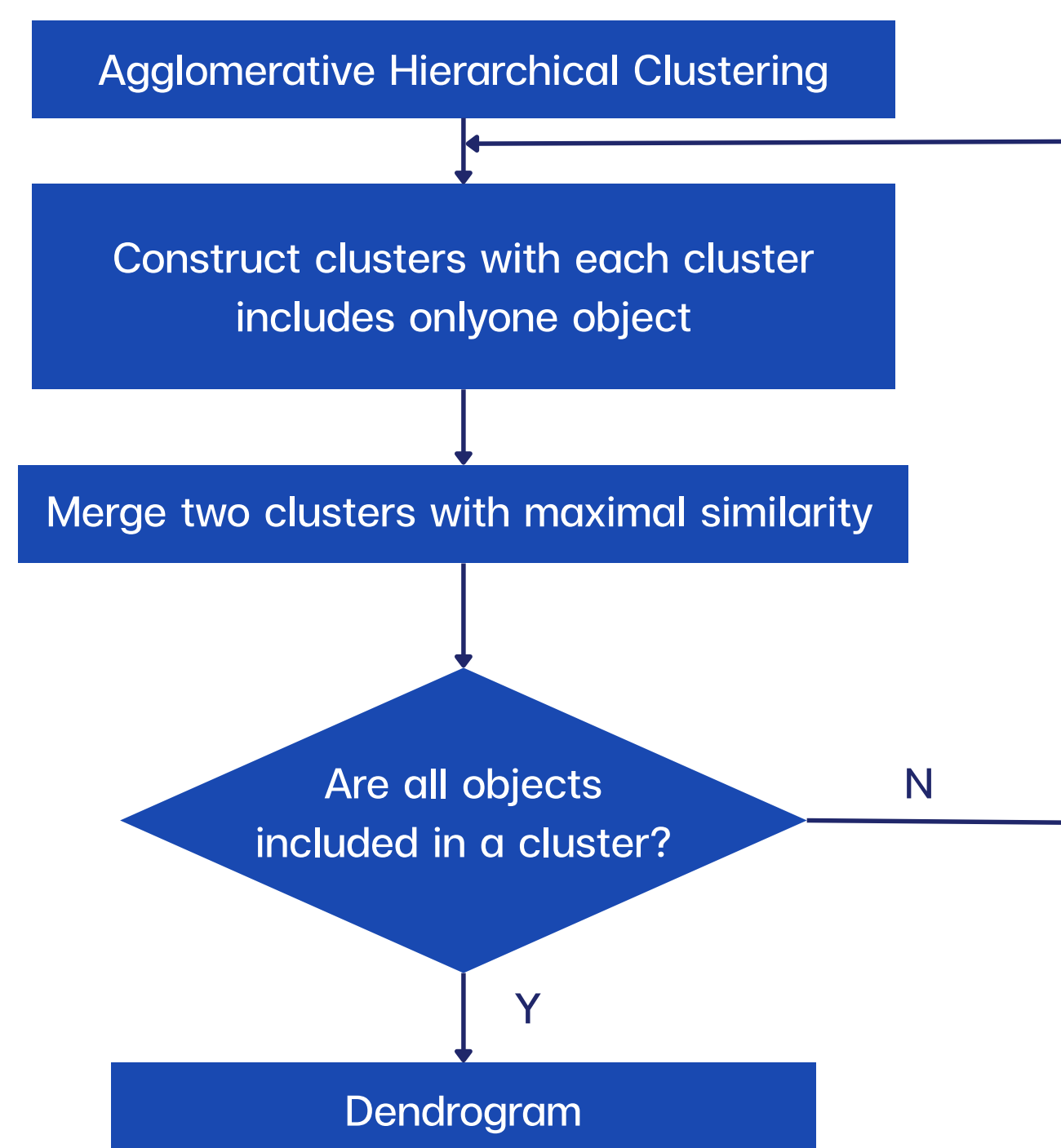


Fig 2. The flowchart of agglomerative hierarchical clustering

## Result and Discussion

The results show similar dendrogram patterns across linkage methods. Ward linkage performs best by clearly separating clusters. In contrast, Single linkage shows a chaining effect, producing poorly defined clusters. Overall, Ward and Complete linkage provide more stable and clear results, while Centroid and Median tend to create uneven and less reliable clusters.

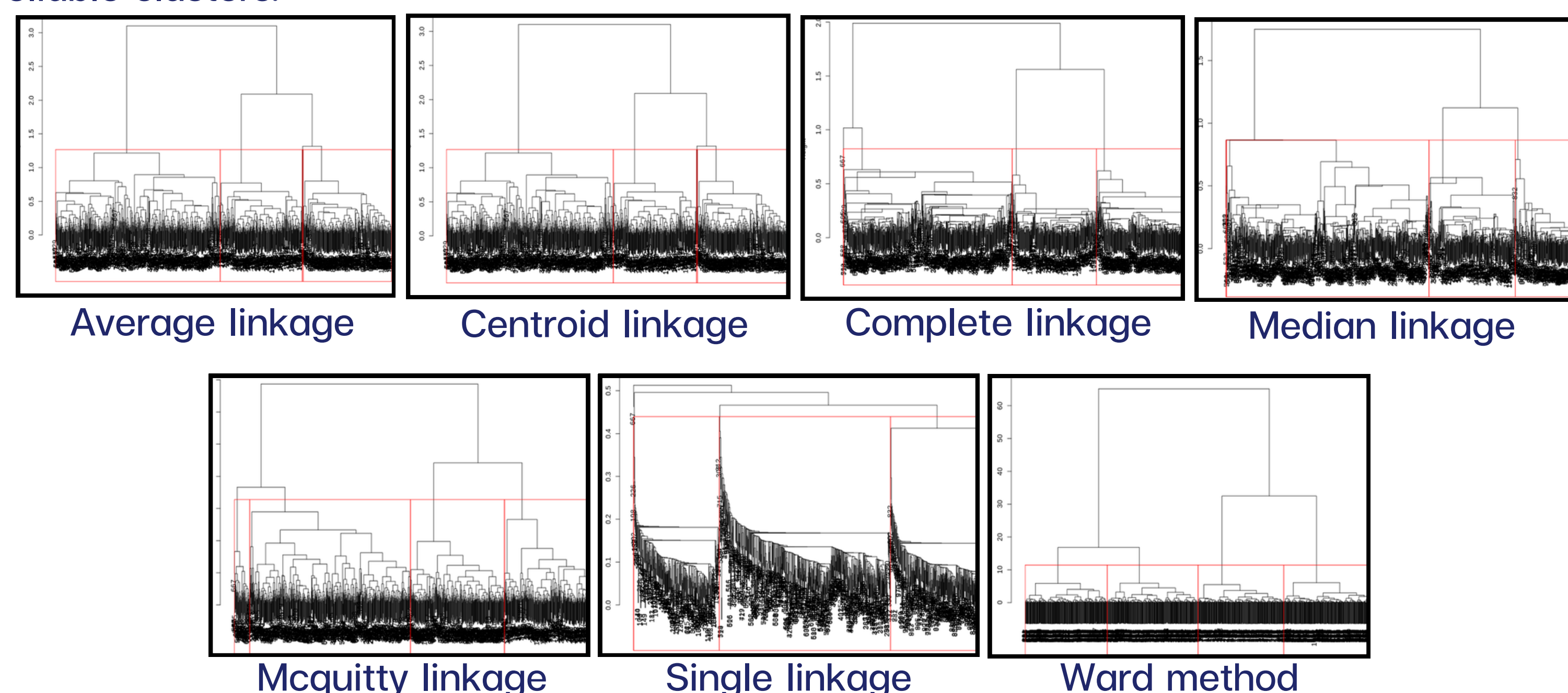


Fig 3. Dendrograms of hierarchical clustering methods

Table 2 Overall performance comparison of linkage methods for simulated D1-D5 and Wine

Dataset	Performance	Linkage						
		Average	Centroid	Complete	Median	McQuitty	Single	Ward
D1	Accuracy	0.762	0.766	0.925	0.678	0.805	0.766	0.986
	Recall	0.6301	0.6304	0.75	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	0.638	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	0.746	0.8041	0.2875	0.7479	0.2511	0.9867
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	0.252	0.7421	0.251	0.9746
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
D5	Accuracy	0.2811	0.2831	0.8807	0.2851	0.7159	0.2831	0.95055
	Recall	0.3204	0.3207	0.9028	0.571	0.8501	0.3207	0.9539
	Precision	0.2477	0.2492	0.871	0.2512	0.7467	0.2493	0.9521
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	0.621	0.3232	0.9495
	Precision	0.7956	0.1255	0.8005	0.1255	0.756	0.1255	0.9407
Wine	Accuracy	0.7284	0.1808	0.7366	0.1808	0.6819	0.1808	0.9451
	Recall							
	Precision							

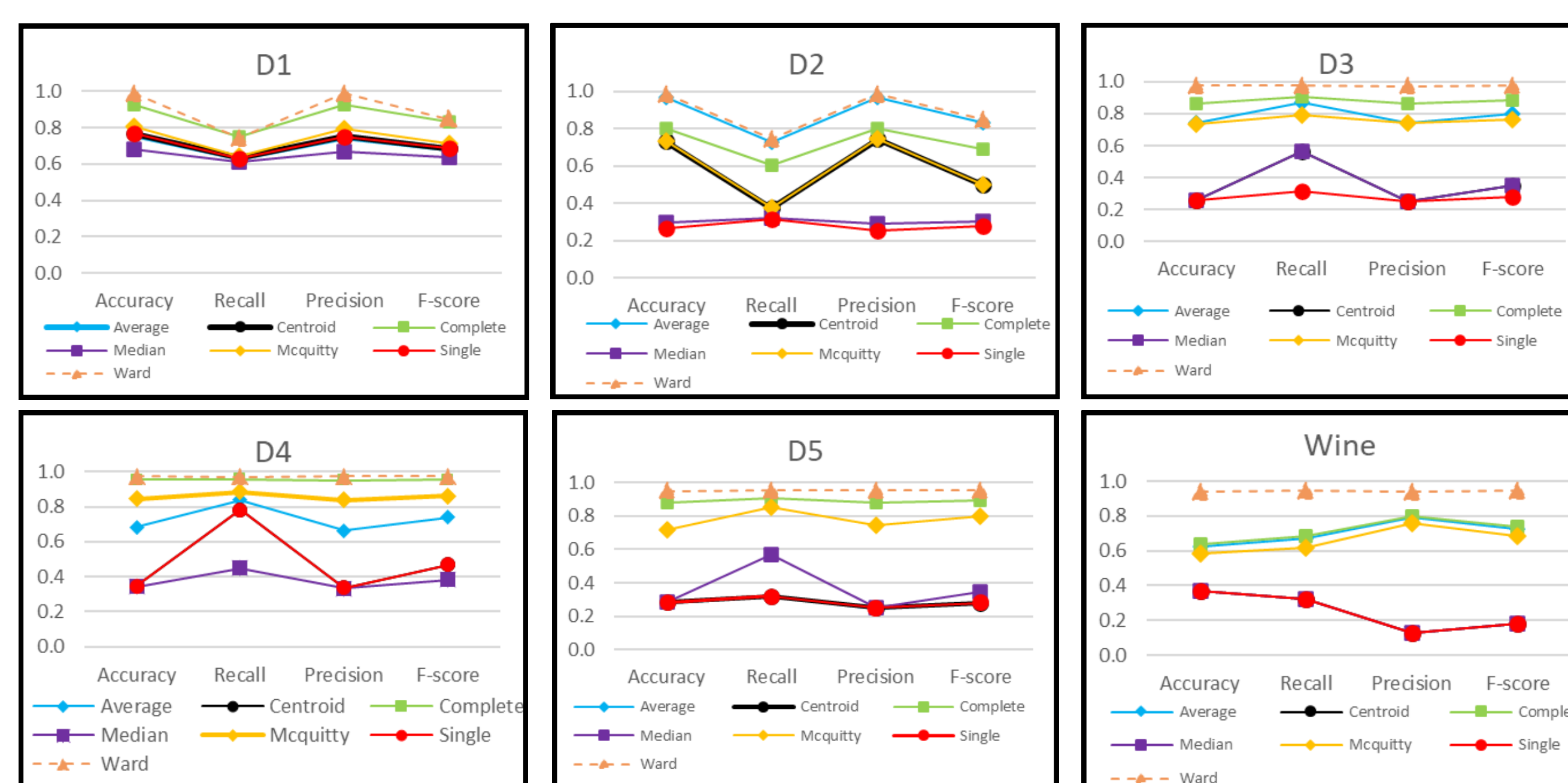


Fig 4. Performance comparison of linkage methods

The graphs show that Ward performs best across both datasets, while Single and Median have the lowest performance. Complete and Average are moderate, suggesting Ward is the most reliable method for this clustering task.

## Conclusion

This study compares hierarchical clustering linkage methods across the Wine and simulated datasets using Accuracy, Precision, Recall, and F-score. Ward performs best and most consistently, while Complete, Average, and McQuitty are moderate. Centroid and Median are less reliable, and Single performs worst due to the chaining effect. Overall, Ward is the most suitable method while others are better used for comparison depending on the data.

## Reference

- IBM.(n.d.).*What is hierarchical clustering?* IBM Think. <https://www.ibm.com/think/topics/hierarchical-clustering>
- van Eck, N. J., & Waltman, L. (2011). Citation-based clustering of publications. arXiv. <https://arxiv.org/pdf/1105.0121>
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
- Sathyanarayanan, S., & Tantri, S. (2024). Confusion matrix-based performance evaluation metrics. African Journal of Biomedical Research.

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.