TEKNIKA: JURNAL SAINS DAN TEKNOLOGI

Homepage jurnal: http://jurnal.untirta.ac.id/index.php/ju-tek/

Hierarchical clustering algorithm-dendogram using Euclidean and Manhattan distance

***Mukhtar Mukhtar*[a,1], *Majid Khan Majahar Ali*[b], *Faula Arina*[a], *Agung Satrio Wicaksono*[a], *Aulia Ikhsan*[a], *Weksi Budiaji*[a], *Syarif Abdullah*[a], *Dinda Dwi Anugrah Pertiwi*[a], *Robby Zidny*[c], *Yuvita Oktarisa*[d], *Royan Habibie Sukarna*[e]**

[a]*Department of Statistics, Faculty of Engineering, University of Sultan Ageng Tirtayasa, Cilegon, Banten 4335, Indonesia*

[b]*School of Mathematics Science, Universiti Sains Malaysia, Gelugor, Pulau Pinang, 11800, Malaysia*

[c]*Department of Chemistry Education, Faculty of Teacher Training and Education, University of Sultan Ageng Tirtayasa, Banten 45583, Indonesia*

[d]*Department of Physic Education, Faculty of Teacher Training and Education, University of Sultan Ageng Tirtayasa, Banten 45583, Indonesia*

[e]*Department of Informatics, Faculty of Engineering, University Sultan Ageng Tirtayasa, Cilegon, Banten, 42435, Indonesia*

[1]Corresponding author: mukhtar@untirta.ac.id

ARTICLE INFO

Article history:
Submitted 06 December 2023
Received 14 December 2023
Received in revised form 20 December 2023
Accepted 24 February 2024
Available online on 18 June 2024

Keywords:
Hierarchical Clustering Algorithm (HCA), Dendogram, Euclidean and Manhattan distance.

Kata kunci:
Hierarchical Clustering Algorithm (HCA), Dendogram, Jarak Euclidean, and Jarak Manhattan.

ABSTRACT

This paper presents the outcomes of a research experiment on the drying process of seaweed. There are numerous approaches to clustering data, such as partitioning and the Hierarchical Clustering Algorithm (HCA). The HCA has been implemented in binary tree structures to visualize data clustering. We conducted a comparative analysis of the four primary methodologies utilized in HCA, namely: 1) single linkage, 2) complete linkage, 3) average linkage, and 4) Ward's linkage. Clustering validation is widely recognized as a crucial issue that significantly impacts the effectiveness of clustering algorithms. Clustering validation can be identified, such as internal and external validation. Internal clustering validation, in particular, holds significant importance in the realm of data science. With this article, the main goal is to do an empirical evaluation of the traits that a representative set of internal clustering validation indices, namely Connectivity, Dunn, and Silhouette, show. In this paper, the HCA applies two distance functions between Euclidean and Manhattan distances to analyze the entanglement function and internal validity.

ABSTRAK

Makalah ini menyajikan hasil percobaan penelitian proses pengeringan rumput laut. Ada banyak pendekatan untuk mengelompokkan data seperti partitioning dan hierarchical clustering algorithm (HCA). HCA telah diterapkan dalam struktur pohon biner untuk memvisualisasikan pengelompokan data. Kami melakukan analisis komparatif terhadap empat metodologi utama yang digunakan dalam HCA yaitu: 1) linkage tunggal, 2) linkage lengkap, 3) linkage rata-rata, dan 4) linkage Ward. Validasi pengelompokan diakui secara luas sebagai masalah penting yang berdampak signifikan terhadap efektivitas algoritma pengelompokan. Validasi clustering dapat diidentifikasi seperti validasi internal dan eksternal. Validasi pengelompokan internal, khususnya, memiliki arti penting dalam bidang ilmu data. Tujuan utama artikel ini adalah untuk melakukan evaluasi empiris terhadap karakteristik yang ditunjukkan oleh kumpulan indeks validasi pengelompokan internal yang representatif, khususnya Konektivitas, Dunn, dan Silhouette. Dalam makalah ini, HCA menerapkan dua fungsi jarak antara jarak Euclidean dan Manhattan untuk menganalisis fungsi keterikatan dan validitas internal.

Available online at http://dx.doi.org/10.62870/tjst.v20i1.23187

1. Introduction

The clustering is the classification of data into groups or clusters [1]. It is the most significant issues in unsupervised learning. It classifies data without label (class) [2]. It is a widely utilized operation in numerous application domains, including exploratory data science and engineering. The process of clustering involves the assignment of each individual object to one or more distinct groups based on certain criteria or characteristics objects in the same group are very similar (intra-cluster similarity/compactness) while objects in different groups are dissimilar (inter-cluster similarity/separation) [3], [4].

The clustering is divided into the following categories: Partitioning and Hierarchical Clustering Algorithms (HCA). The research only emphasizes on HCA approach which narrow down to agglomerative [5]. In a HCA, due to the multiple resolutions of the clusters, it is possible to recursively divide a sizable cluster into smaller sub-clusters [6].

The HCA can be classified as either agglomerative, also known as "bottom-up," or divisive, also known as "top-down". Agglomerative algorithms initiate the clustering process by considering each element as an individual cluster. Subsequently, these clusters are progressively merged together to form larger clusters [7], [8].

The HCA is a data analysis technique that involves the grouping of data objects into a hierarchical tree-like structure known as a cluster. It generates a nested sequence of partitions, with a single, all-inclusive cluster at the top and singletons of individual objects at the bottom. The concept of an intermediate level can be viewed as the combination of two clusters from the previous lower level or the division of a cluster from the subsequent higher level. The graphical representation of the output from a HCA is commonly depicted as a dendrogram, which visually resembles a tree structure. The merging process and intermediate clusters are depicted graphically in this tree. The visual representation depicts the process of combining points into a solitary cluster [9].

The dendrogram is a useful tool to visualize the outcomes of HCA. It visually representations that depict the hierarchical relationships between entities based on their levels of dissimilarity and similarity. On the right side of the dendrogram, every individual observation is as an independent cluster. For each observation, horizontal lines proceed up at different values between "dissimilarity" and "similarity", these lines have connections to lines generated by other observations using lines that are horizontal. The procedure of observation continues until all of the observations are clustered together on the right side of the dendogram [10].

In the context of clustering, distance is a crucial parameter to identify clusters. Distance measures can be utilized to calculate the degree of similarity between objects [11]. The aims of the research are to explore different distance measures that could be applied in this clustering and to evaluate how different distance measures in HCA such as single, complete, average, and Ward's linkage method would affect the clustering output. The distance measures applied in this research includes Euclidean and Manhattan distance.

2. Methodology

2.1. Hierarchical Clustering Algorithm

The HCA is utilized to arrange data in a hierarchical structure based on the proximity matrix. Linkage is a metric used to assess the proximity between two distinct clusters of elements. There are different of linkages namely single, complete, average, and wards.

Table 1. Hierarchical Clustering Algorithm

Method's	Distance update formula for $d(I \cup J, K)$	Cluster dissimilarity between clusters A and B
Single	$\min(d(I,J), d(J,K))$	$min_{a\in A, b\in B} d[a,b]$
Complete	$\max(d(I,J), d(J,K))$	$max_{a\in A, b\in B} d[a,b]$
Average	$\frac{n_I d(I,K) + n_J d(J,K)}{n_I + n_J}$	$\frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{a\in A}\sum_{b\in B} d[a,b]$
Ward	$\frac{n_i+n_k}{n_i+n_j+n_k} d(C_i,C_k) + \frac{n_j+n_k}{n_i+n_j+n_k} d(C_j,C_k) - \frac{n_i}{n_i+n_j+n_k} d(C_i,C_j)$	$\sqrt{\frac{2\lvert A\rvert\lvert B\rvert}{\lvert A\rvert+\lvert B\rvert}} \cdot \lVert \vec{C}_A - \vec{C}_B \rVert_2$

Clustering based on single linkage, the single element pair, specifically those two elements (one in each cluster) that are located in the closest proximity to each other, is used to calculate the distance that separates two clusters. This algorithm is also referred to as nearest neighbor clustering [12]. The complete linkage algorithm defines inter-cluster distance using the farthest distance between two objects [13]. The Ward's linkage can be achieved through the utilization of the Lance-Williams formula [14]. The Average linkage is the average distance between elements within each cluster. The distance between any two clusters A and B, each of size (i.e., cardinality) $|A|$ and $|B|$, is taken to be the average of all distances $d(x, y)$ between pairs of objects x in A and y in B [15].

2.2. Euclidean Distance

When presented with two instances in p-dimensions, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$, The calculation of the distance between two data instances can be performed using the Minkowski metric [16].

$$d(x_i, x_j) = \left(\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \left|x_{i3} - x_{j3}\right|^2 + \dots + \left|x_{ip} - x_{jp}\right|^2\right)^{\frac{1}{2}} \quad (1)$$

$$d(x_i, x_j) = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_{i1} - x_{j1}\right|^2\right)^{\frac{1}{2}} \quad (2)$$

2.3. Manhattan Distance

Manhattan distance between two items is the sum of their component differences [17]. The distance between a point $x = (x_1, x_2, \dots x_n)$ and a point $y = (y_1, y_2, \dots y_n)$ is:

$$MD_{(x,y)} = \sum_{i=1}^{n} |x_i - y_i|, \tag{3}$$

where the variables x_i and y_i represent the values of the i^{th} variable at points x and y, respectively, with n denoting the number of variables.

2.4. Connectivity

The concept of measuring connectivity is derived from graph theory [18]. Specify as $nn_{i(j)}$ the j th shortest neighbor of observation i, and let $x_{i,nn_{i(j)}}$ be zero if i and $nn_{i(j)}$ are in the same cluster and $1/j$ otherwise. Then, for a specific clustering partition $C = \{C_i, \dots, C_k\}$ of the N observations into K disjoint clusters, the definition of connectivity is

$$\mathrm{Conn(C)} = \sum_{i=1}^{M} \sum_{j=1}^{L} x_{i,nn_{i(j)}}. \tag{4}$$

Connection values range from 0 to infinity (∞) and should be minimized [19].

2.5. Dunn Indexed

Dunn's index ought to be maximized [20]. The range of the Dunn index is zero (0) to infinity (∞). The formula for the Dunn index is

$$\mathrm{DI} = \frac{d_{min}}{d_{max}} \tag{5}$$

$$d_{min} = min\{d(x,y);\ x \in C_i, y \in C_{j,} i \neq j\} \tag{6}$$

$$d_{max} = max\{d(x,y);\ x \in C_i, y \in C_{j,} i = j\} \tag{7}$$

2.6. Silhouette

The silhouette value expresses the degree of certainty in the clustering assignment of a specific observation, with values close to 1 (positive) for well-clustered observations and Unwell clustered observations with values close to -1 (negative) [21]. The definition of silhouette for observation i is:

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}}, \tag{8}$$

$$b_i = min \frac{\sum_j d(i,j)}{|c_j|}, C_i \neq C_j, \text{ and} \tag{9}$$

$$a_i = min \frac{\sum_j d(i,j)}{|c_j|}, C_i = C_j, \tag{10}$$

where, a_i is the average distance between observation i and all other observations in the same cluster and b_i average distance in the closest neighboring cluster between observation i and all other observations.

3. Result and Discussion

3.1 Hierarchical Clustering Algorithm of Euclidean Distance

The central issue is determining the value of the parameter k (cluster). Furthermore, the second difference and D-index index D R package for determining the quantity of clusters. The following four images are provided for the purpose of determining the number of clusters through hierarchical analysis.

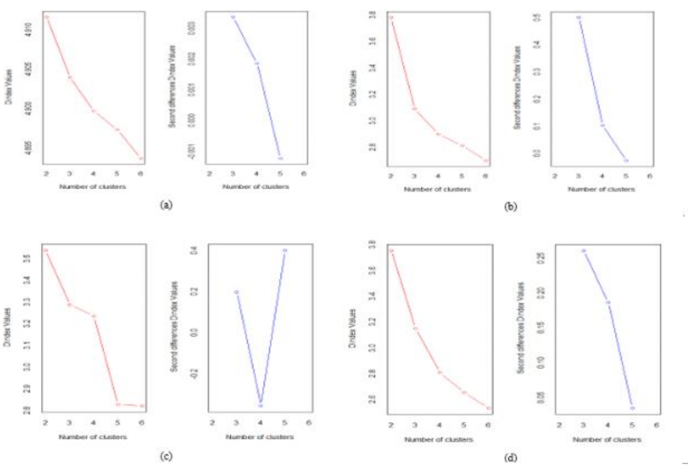

Figure 1: Number of Cluster:
(a) Single, (b) Complete, (c) Average, and (d) Ward-Linkage

It is important to emphasize that this approach consistently considers the majority of the indexes pertaining to each cluster size. The best number of clusters is 3, which is easily visible in the second differences D-index graph. The Euclid distance has been employed to determine the distance between the data. This study constructed dendrograms resulting from cluster analysis for objective functions among single, complete, average, and ward method in order to discuss the results using the Euclidean distance.

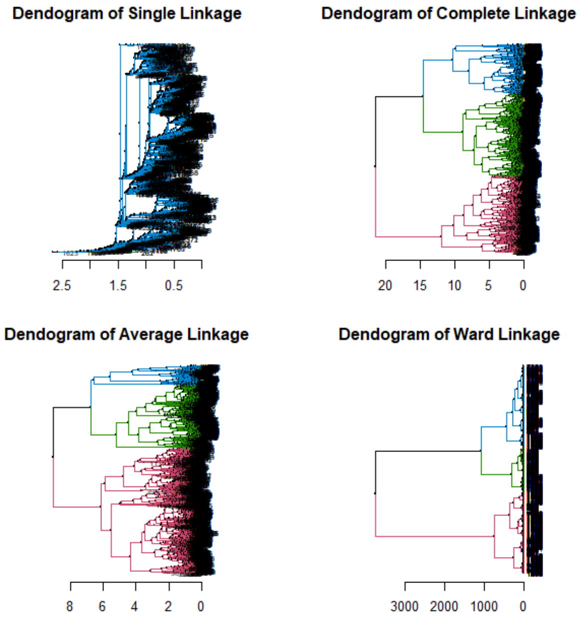

Figure 2: Dendrograms for Euclid Distance

Four dendogram images are displayed above. The results examine the quantity of each subcluster within every dendogram.

Table 2. Sub-cluster from Dendogram

Hierarchical	Sub-cluster	Value
Single	Cluster 1	1911
	Cluster 2	2
	Cluster 3	1
Complete	Cluster 1	453
	Cluster 2	758
	Cluster 3	703
Average	Cluster 1	175
	Cluster 2	581
	Cluster 3	1158
Ward	Cluster 1	412
	Cluster 2	1097
	Cluster 3	405

The member counts for each cluster are presented in Table 2, categorized by ward, single, complete, average, and ward. For single-linkage cluster 1 had majority members which 1911 members. Complete-linkage cluster 2 had majority members which 758 members. Average-linkage cluster 3 had majority members which 1158 members. Ward-linkage cluster 2 had majority 1097.

The internal validation of clusters is of utmost importance in the field of clustering. In this analysis, the result objectively discusses various techniques of cluster validation.

Table 3. Internal Validation for Euclid Distance

Hierarchical	Internal Validation	Value
Single	Connectivity	6.7869
	Dunn	0.1201
	Silhouette	0.0551
Complete	Connectivity	166.3413
	Dunn	0.0349
	Silhouette	0.3045
Average	Connectivity	75.9365
	Dunn	0.0341
	Silhouette	0.3361

Hierarchical	Internal Validation	Value
	Connectivity	111.5837
Ward	Dunn	0.0550
	Silhouette	0.2962

According to the findings presented in Table 3, the connectivity value (minimized) is recorded as 6.7869, specifically observed under the single-linkage method. The Dunn index achieves a minimum value of 0.0341 when utilizing the average-linkage method. The maximum value of the silhouette at average linkage is 0.3361.

3.2 Hierarchical Clustering Algorithm of Manhattan Distance

Figure 3 illustrates four images that are used to determine the number of clusters in each hierarchical method, specifically using the Manhattan distance metric.

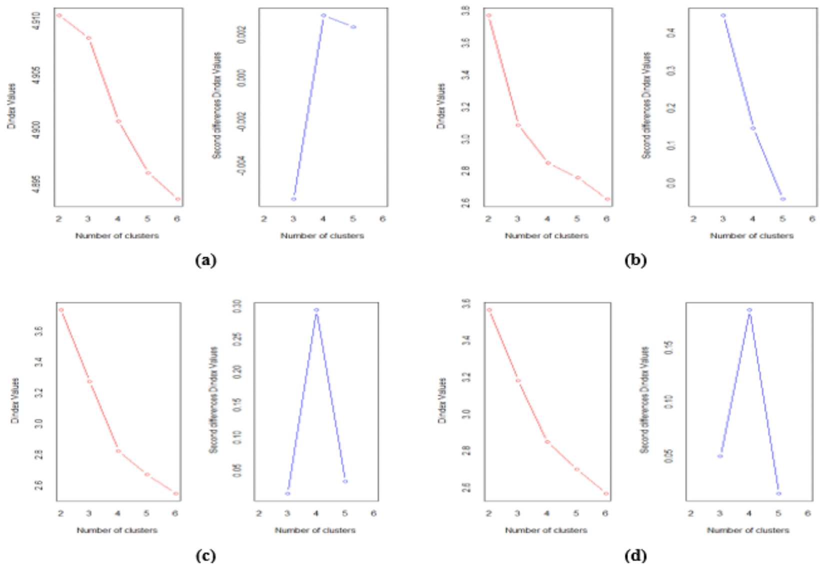

Figure 3. Number of Cluster:
(a) Single, (b) Complete, (c) Average, and (d) Ward-Linkage

Figure 3 shown the best number of clusters is 2 for Single-Linkage, Complete-Linkage is 3 clusters, Average-Linkage is 3 clusters, and Ward-Linkage is 2 clusters. Figure 4 is four images of dendograms. The results examine the quantity of every subgroup within every dendogram.

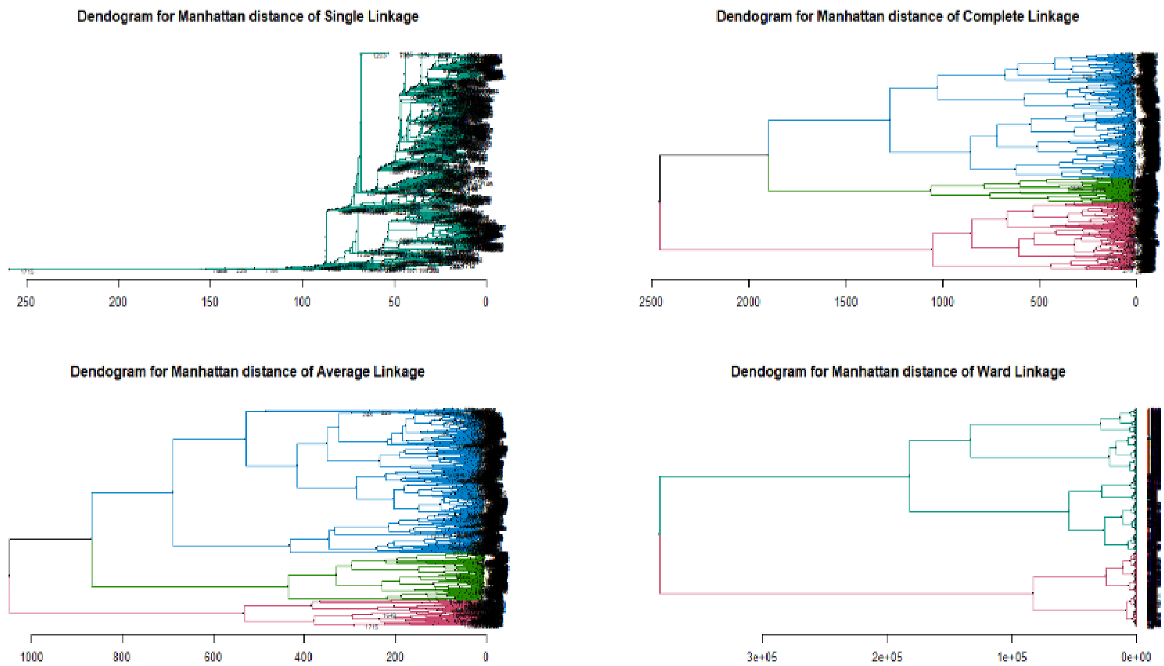

Figure 4. Dendrograms for Manhattan Distance

Table 4. Sub-cluster from Dendogram

Hierarchical	Sub-cluster	Value
Single	Cluster 1	1913
	Cluster 2	1
Complete	Cluster 1	462
	Cluster 2	735
	Cluster 3	717
Average	Cluster 1	419
	Cluster 2	1283
	Cluster 3	212
Ward	Cluster 1	417
	Cluster 2	1497

Table 4 shows the number of members for each cluster when applied amongst single, complete, average, and wards. For single-linkage cluster 1 had majority members which 1913 members. Complete-linkage cluster 2 had majority members which 735 members. Average-linkage cluster 2 had majority members which 1283 members. Ward-linkage cluster 2 had majority 1497.

Table 5. Internal Validation for Manhattan Distance

Hierarchical	Internal Validation	Value
Single	Connectivity	2.9290
	Dunn	0.0806
	Silhouette	0.1654
Complete	Connectivity	138.6984
	Dunn	0.0306
	Silhouette	0.3624
Average	Connectivity	45.1841
	Dunn	0.0233
	Silhouette	0.3566
Ward	Connectivity	25.6734
	Dunn	0.0200
	Silhouette	0.4783

It is also essential to validate clusters objectively discuss several techniques of cluster validation. Table 5 shows amongst the connectivity value (minimize) is 2.9290 at single-linkage. The Dunn value (minimize) is 0.0200 at Ward’s-linkage. The Silhouette value (maximize) is 0.4783 at Ward’s-linkage. From table above that Ward-linkage method better than others.

4. Conclusion

This paper investigated the use of Euclidean distances and Manhattan distance amongst Single, complete, average, and Ward’s-linkage method. And comparing entanglement function each other’s. For Euclid distance between average and complete entanglement value which has a very high similarity is 0.33. The entanglement average versus ward's has many differences is 0.91. Validity shown amongst the connectivity value (minimize) is 6.7869 at single-linkage. The Dunn value (minimize) is 0.0341 at average-linkage. The Silhouette value (maximize) is 0.3361 at average-linkage. The values above that Average-linkage method better than others. For Manhattan distance between complete versus ward’s entanglement value which has a very high similarity is 0.33. The entanglement average versus ward's has many differences is 0.84. Validity shown amongst the connectivity value (minimize) is 2.9290 at single-linkage. The Dunn value (minimize) is 0.0200 at Ward’s-linkage. The Silhouette value (maximize) is 0.4783 at Ward’s-linkage. From table above that Ward-linkage method better than others. In future, the research may be extended by considering for dendogram between normal and un-normal data to improve the clustering accuracy.

References

[1] Abbas, K. A. *et al.* (2023). Unsupervised machine learning technique for classifying production zones in unconventional reservoirs. *Int. J. Intell. Networks*, vol. 4, pp. 29–37. Doi: 10.1016/j.ijin.2022.11.007.

[2] Huang, J., Yu, Z. L., & Gu, Z. (2018). A clustering method based on extreme learning machine. *Neurocomputing*, vol. 277, pp. 108–119, Feb. 2018. Doi: 10.1016/j.neucom.2017.02.100.

[3] Chhabra, A., Masalkovaite, K., & Mohapatra, P. (2021). An Overview of Fairness in Clustering. *IEEE Access*, vol. 9, pp. 130698–130720. Doi: 10.1109/ACCESS.2021.3114099.

[4] Thilagavathi, G., Srivaishnavi, D., & Aparna, N. (2013). A Survey on Efficient Hierarchical Algorithm used in Clustering. *Int. J. Eng. Res. Technol.*, vol. 2, no. 9, pp. 2553–2556.

[5] Saket, S., & Pandya, S. (2016). Implementation of Extended K-Medoids Algorithm to Increase Efficiency and Scalability using Large Datasets. *Int. J. Comput. Appl.*, vol. 146, no. 5, pp. 19–23, Jul. 2016. Doi: 10.5120/ijca2016910701.

[6] Krishnamurthy, L. *et al.* (2011). Large genetic variation for heat tolerance in the reference collection of chickpea (Cicer arietinum L.) germplasm.

Plant Genet. Resour., vol. 9, no. 01, pp. 59–69, Apr. 2011. Doi: 10.1017/S1479262110000407.

[7] Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview II. *WIREs Data Min. Knowl. Discov.*, vol. 7, no. 6, Nov. 2017. Doi: 10.1002/widm.1219.

[8] Zhang, Z., Murtagh, F., Van Poucke, S., Lin, S., & Lan, P. (2017). Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. *Ann. Transl. Med.*, vol. 5, no. 4, pp. 75–75, Feb. 2017. Doi: 10.21037/atm.2017.02.05.

[9] Rani, Y., & Rohil, H. (2013). A Study of Hierarchical Clustering Algorithm. *International Journal of Information and Computation Technology*, vol. 3, no. 11, pp. 1225–1232.

[10] Sembiring, R. W., Zain, J. M., & Embong, A. (2011). A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course. No. January, 2011, [Online]. Available: http://arxiv.org/abs/1101.4270

[11] Miller, H. J. (2007). Geographic Data Mining and Knowledge Discovery, in *The Handbook of Geographic Information Science*, Wiley, 2007, pp. 352–366. doi: 10.1002/9780470690819.ch19.

[12] Camiz, S. & Pillar, V. (2007). Comparison of single and complete linkage clustering with the hierarchical factor classification of variables. *Community Ecol.*, vol. 8, no. 1, pp. 25–30, Jun. 2007. Doi: 10.1556/ComEc.8.2007.1.4.

[13] Gere, A. (2023). Current Research in Food Science Recommendations for validating hierarchical clustering in consumer sensory projects. *Curr. Res. Food Sci.*, vol. 6, no. May, p. 100522, 2023. Doi: 10.1016/j.crfs.2023.100522.

[14] Murtagh, F. (2014). Ward ' s Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward ' s Criterion ?. *J. Classif.*, vol. 295, no. October, pp. 274–295. Doi: 10.1007/s00357-.

[15] Xu, N., Finkelman, R. B., Dai, S., Xu, C., & Peng, M. (2021). Average Linkage Hierarchical Clustering Algorithm for Determining the Relationships between Elements in Coal. *ACS Omega*, vol. 6, no. 9, pp. 6206–6217, Mar. 2021. Doi: 10.1021/acsomega.0c05758.

[16] Han & Kamber. (2011). *Data Mining: Concepts and Techniques*, 3rd Editio. Burlington: Morgan Kaufmann.

[17] Ponnmoli & Selvamuthukumaran. (2014). Analysis of face recognition using manhattan distance algorithm with image segmentation. *Intl. J. Comput. Sci. Mob. Comput*, vol. 3, pp. 18–27.

[18] Vergani, A. A. & Binaghi, E. (2018). A soft davies-bouldin separation measure. *IEEE Int. Conf. Fuzzy Syst.*, vol. 2018-July, no. February, 2018. Doi: 10.1109/FUZZ-IEEE.2018.8491581.

[19] Yahyaoui, H., & Own, H. S. (2018). Unsupervised clustering of service performance behaviors. *Inf. Sci. (Ny).*, vol. 422, pp. 558–571, Jan. 2018. Doi: 10.1016/j.ins.2017.08.065.

[20] Pal, N. R., & Biswas, J. (1997). Cluster validation using graph theoretic concepts," *Pattern Recognit.*, vol. 30, no. 6, pp. 847–857. Doi: 10.1016/S0031-3203(96)00127-6.

[21] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65. Doi: 10.1016/0377-0427(87)90125-7.