Experimental study of the behaviour of the Calinski-Harabasz index across the α-cuts of a hierarchical clustering dendrogram and its suitability for automatic dendrogram cutting

Sergi Ramirez-Mitjans[1,2], Dante Conti[1,2], and Karina Gibert[1,2]

[1] Universitat Politècnica de Catalunya, Barcelona, Spain
[2] Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence (IDEAI-UPC), Barcelona, Spain

Abstract. Determining the most appropriate cut of a hierarchical clustering dendrogram remains one of the main challenges in cluster analysis. Although the Calinski–Harabasz (CH) index is one of the most widely used cluster validation criteria, its behaviour across the complete sequence of dendrogram α-cuts has received limited attention in the literature.

This paper presents an experimental study of the behaviour of the Calinski–Harabasz index across the partitions generated by hierarchical clustering dendrograms and evaluates its suitability as an automatic dendrogram cutting criterion. Three synthetic datasets were generated under controlled conditions representing well-separated clusters, overlapping clusters and noisy cluster structures. For each scenario, the evolution of the CH index was analysed across all possible dendrogram cuts and compared with the known ground-truth clustering structure.

The results show that the Calinski–Harabasz index successfully identifies the true number of clusters when the underlying groups are clearly separated and remains robust under moderate levels of noise contamination. In contrast, cluster overlap substantially affects the behaviour of the criterion, producing flatter CH profiles and leading to an overestimation of the number of clusters. The experiments also suggest that the overall shape of the CH curve provides valuable information regarding the reliability of the selected partition beyond the location of its maximum value.

These findings support the use of the Calinski–Harabasz index as an automatic dendrogram cutting criterion in hierarchical clustering while highlighting the importance of analysing its behaviour across the complete hierarchy.

Keywords: Hierarchical clustering · Dendrogram cutting · Calinski–Harabasz index · Cluster validation · α-cuts · Automatic tree cutting

1 Introduction

Hierarchical clustering is one of the most widely used unsupervised learning techniques for exploratory data analysis and knowledge discovery. Unlike parti-

tioning methods, hierarchical clustering produces a complete hierarchy of nested partitions represented through a dendrogram, allowing analysts to explore the structure of the data at different levels of granularity.

One of the most important challenges associated with hierarchical clustering is determining the most appropriate cut of the dendrogram. While the clustering process itself does not require specifying the number of clusters in advance, the final partition depends on selecting a suitable level within the hierarchical structure. Consequently, automatic dendrogram cutting remains an active research topic in cluster analysis.

Several approaches have been proposed to address this problem. Some methods rely on visual inspection of dendrograms, whereas others use cluster validation criteria to identify meaningful partitions automatically. Among the latter, the Calinski–Harabasz (CH) index [1] is one of the most widely used due to its computational simplicity and its ability to simultaneously evaluate cluster compactness and separation.

The original work of Calinski and Harabasz proposed a variance-ratio criterion for evaluating consecutive hierarchical partitions. More specifically, the method was formulated as an iterative statistical procedure comparing neighbouring cuts of a hierarchical tree [1]. However, its application requires several distributional assumptions that are often difficult to guarantee in practical clustering problems, particularly when dealing with noisy, heterogeneous or mixed-type datasets.

In recent years, several studies have investigated the use of the Calinski–Harabasz index as an automatic dendrogram cutting criterion. Particularly relevant are the contributions by Karna and Gibert [15], who analysed different strategies derived from the CH index for identifying meaningful partitions in hierarchical clustering. Their results showed that the behaviour of the index across consecutive dendrogram cuts contains valuable information for supporting automatic tree-cutting decisions.

Despite these advances, the behaviour of the Calinski–Harabasz index across the complete sequence of dendrogram α-cuts remains insufficiently studied. In particular, there is limited experimental evidence regarding its robustness under different cluster configurations, separation levels and noise conditions.

The objective of this paper is to experimentally analyse the behaviour of the Calinski–Harabasz index across the sequence of α-cuts generated by hierarchical clustering dendrograms and to evaluate its suitability as an automatic dendrogram cutting criterion.

To this end, several synthetic datasets were generated under controlled conditions representing three common clustering scenarios: well-separated clusters, overlapping clusters and noisy cluster structures. For each scenario, the evolution of the Calinski–Harabasz index was analysed across all dendrogram cuts and compared with the known underlying cluster structure.

The results provide empirical evidence regarding the strengths and limitations of the Calinski–Harabasz criterion for automatic dendrogram cutting and

contribute to a better understanding of its behaviour in hierarchical clustering applications.

2 Related Work

Several cluster validation indices have been proposed for determining appropriate dendrogram cuts, including the Dunn index [5], Davies–Bouldin index [4], Silhouette coefficient [3] and Calinski–Harabasz (CH) index [1].

The original CH criterion was formulated in the context of hierarchical clustering and evaluates the trade-off between cluster separation and compactness. Previous studies have shown its effectiveness as an internal validation criterion [6].

More recently, Karna and Gibert [15] analysed automatic dendrogram cutting strategies derived from the CH index and showed that the behaviour of the criterion across consecutive dendrogram cuts contains valuable information for identifying meaningful partitions.

The present work extends this line of research by experimentally analysing the behaviour of the CH index across the complete sequence of dendrogram α-cuts under different cluster separation and noise conditions.

3 Hierarchical Clustering and Dendrogram α-Cuts

Hierarchical clustering produces a dendrogram representing a sequence of nested partitions. Different horizontal cuts of the dendrogram generate different cluster solutions, commonly referred to as α-cuts (that are obtained when the height of the dendogram changes).

For a dataset containing n observations, an agglomerative hierarchical clustering process generates a sequence of partitions ranging from n singleton clusters to a single cluster containing all observations.

For each partition, the Calinski–Harabasz index can be computed as

$$CH(k) = \frac{BSS(k)/(k-1)}{WSS(k)/(n-k)}, \tag{1}$$

where $BSS(k)$ and $WSS(k)$ denote the between-cluster and within-cluster sums of squares, respectively.

4 Experimental Design

The objective of the experimental study is to analyse the behaviour of the Calinski–Harabasz index across the sequence of α-cuts generated by hierarchical clustering dendrograms. Particular attention is given to evaluating the suitability of the index as an automatic dendrogram cutting criterion under different structural conditions.

To this end, three synthetic datasets were generated representing common situations encountered in clustering applications: well-separated clusters, overlapping clusters and noisy cluster structures. The use of synthetic data allows complete control over the underlying cluster configuration and provides a known ground truth against which the behaviour of the validation criterion can be assessed.

4.1 Synthetic Data Generation

Three synthetic datasets were generated using two numerical variables (x and y) in order to represent increasing levels of clustering difficulty while maintaining the same underlying number of clusters.

Table 1 summarizes the characteristics of the synthetic datasets considered in the study.

Table 1. Characteristics of the synthetic datasets.

Scenario	Observations	Variables	True Clusters	Noise
Well-separated	240	2	3	No
Overlapping	240	2	3	No
Noisy	255	2	3	Yes

The first dataset represents an ideal clustering scenario and contains three well-separated Gaussian clusters composed of 80 observations each. The second dataset preserves the same structure but reduces the distance between centroids, creating substantial overlap between neighbouring groups. The third dataset contains three clusters of 70 observations each together with 45 randomly generated observations acting as noise.

These datasets were designed to represent three common clustering situations: well-separated groups, overlapping structures and noisy data.

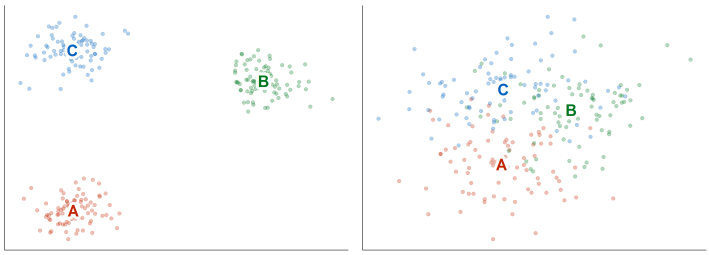
4.2 Hierarchical Clustering Configuration

All experiments were performed using agglomerative hierarchical clustering.

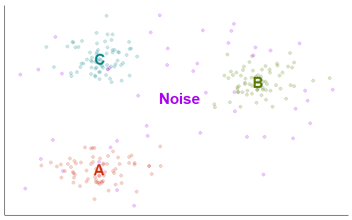
The Euclidean distance was used to compute pairwise dissimilarities between observations. Cluster aggregation was performed using Ward's linkage criterion (Ward.D2), which minimizes the increase in within-cluster variance at each merge operation.

Ward's method was selected because of its widespread use in practical applications and because it is directly related to the variance decomposition underlying the Calinski–Harabasz index.

For each synthetic dataset, a complete dendrogram was generated and all possible α-cuts were analysed.



(a) Well-separated (b) Overlapping



(c) Noisy

Fig. 1. Synthetic datasets used in the experimental study. The three scenarios represent increasing levels of clustering difficulty, ranging from clearly separated clusters to overlapping and noisy structures.

4.3 Validation Indicators

The following indicators were computed for every partition induced by the dendrogram: Number of clusters (k), Within-cluster sum of squares (WSS), Between-cluster sum of squares (BSS), Total sum of squares (TSS), Calinski–Harabasz index (CH) and dendrogram height associated with each merge operation.

The Calinski–Harabasz index was computed for every possible partition generated by the dendrogram. The resulting sequence of CH values was subsequently analysed to identify the partitions preferred by the criterion under each experimental scenario.

In addition to the maximum CH value, the overall shape of the CH curve, the sharpness of its maxima and its relationship with dendrogram height variations were also examined. This analysis provides additional information regarding the reliability of the partitions suggested by the index and facilitates a deeper understanding of its behaviour across different clustering structures.

The relative variation between consecutive dendrogram heights considerated as α-cuts are also plotted in the figures. This is an alternative way for dendrogram cutting, as it corresponds to the largest agglomeration jump in the hierarchy.

5 Results

This section presents the behaviour of the Calinski–Harabasz index across the sequence of α-cuts generated by the hierarchical clustering dendrograms.

For each synthetic scenario, the evolution of the index was analysed together with the corresponding dendrogram structure and the known ground-truth number of clusters. Particular attention was paid to the location of the maximum CH value, the overall shape of the CH curve and the sharpness of the detected maxima.

The objective is to evaluate whether the partitions suggested by the Calinski–Harabasz criterion coincide with the underlying clustering structure and whether the behaviour of the index remains stable under increasingly difficult clustering conditions.

5.1 Well-Separated Clusters

The first scenario represents the most favourable clustering conditions, with three compact and clearly distinguishable groups. Figure 2 presents the evolution of the Calinski–Harabasz index across the sequence of dendrogram α-cuts.

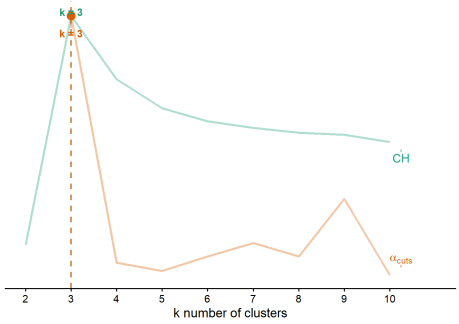


Fig. 2. Evolution of the Calinski–Harabasz index for the well-separated scenario.

The maximum value of the Calinski–Harabasz index was obtained for $k = 3$, coinciding exactly with the true number of clusters used to generate the dataset.

The maximum CH value reached 2534.78, substantially larger than the values observed for neighbouring partitions. This produces a very pronounced peak in the CH curve, making the identification of the optimal partition straightforward.

The results indicate that, when the cluster structure is clearly defined, the Calinski–Harabasz criterion provides an accurate and unambiguous indication of the appropriate dendrogram cut.

Moreover, the corresponding dendrogram exhibits a clear separation between the three groups, with the largest merge occurring after the partition containing three clusters has already been formed. This behaviour is consistent with the interpretation of CH as a measure balancing cluster compactness and separation.

5.2 Overlapping Clusters

The second scenario introduces partial overlap between neighbouring clusters, creating a more challenging clustering problem. Figure 3 shows the evolution of the Calinski–Harabasz index for this dataset.

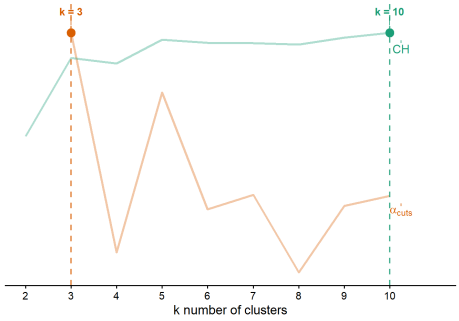


Fig. 3. Evolution of the Calinski–Harabasz index for the overlapping scenario.

Unlike the previous scenario, the maximum value of the Calinski–Harabasz index does not coincide with the true number of clusters.

The largest CH value was obtained for $k = 10$, whereas the data were generated from three underlying groups. Furthermore, the maximum value reached only 171.97, considerably lower than that observed in the well-separated scenario.

An additional characteristic of this case is the absence of a clearly dominant peak. The CH curve presents several local maxima of similar magnitude, producing a much flatter profile.

This behaviour suggests that the index interprets density fluctuations generated by cluster overlap as additional cluster structure. Consequently, the criterion tends to overestimate the number of clusters when the underlying groups are not clearly separated.

The results indicate that the reliability of the Calinski–Harabasz criterion decreases as cluster overlap increases, particularly when cluster boundaries become difficult to distinguish.

5.3 Noisy Clusters

The third scenario evaluates the robustness of the Calinski–Harabasz index in the presence of random noise observations. Figure 4 presents the corresponding CH curve.

Despite the presence of 45 noise observations, the Calinski–Harabasz index correctly identified the true number of clusters.

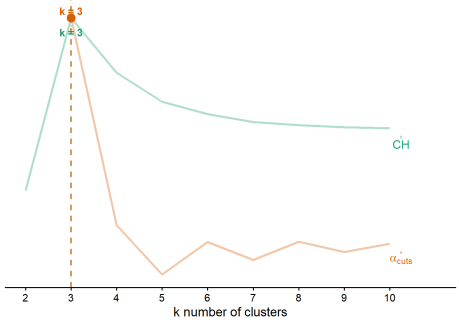


Fig. 4. Evolution of the Calinski–Harabasz index for the noisy scenario.

The maximum value was obtained for $k = 3$, matching the underlying cluster structure used during data generation.

As expected, the absolute value of the index decreased substantially compared with the well-separated scenario, reaching a maximum value of 853.94. Nevertheless, the location of the maximum remained unchanged and the peak continued to be clearly distinguishable from neighbouring partitions.

These results suggest that the Calinski–Harabasz index exhibits a considerable degree of robustness against moderate levels of noise contamination.

Although noise reduces cluster compactness and partially deteriorates cluster separation, the criterion remains capable of identifying the most meaningful dendrogram cut.

5.4 Comparative Analysis

Figure 5 compares the evolution of the Calinski–Harabasz index across the three experimental scenarios. Table 2 summarizes the main results obtained in terms of the true number of clusters, the partition selected by the maximum CH value and the corresponding maximum index value.

Table 2 summarizes the results obtained for all datasets.

Table 2. Summary of experimental results.

Scenario	True k	Selected k	Maximum CH	Correct
Well-separated	3	3	2534.78	Yes
Overlapping	3	10	171.97	No
Noisy	3	3	853.94	Yes

The comparison reveals three clearly differentiated behaviours. In the well-separated scenario, the CH profile presents a sharp and dominant maximum at $k = 3$, which coincides with the true generative structure of the data. This

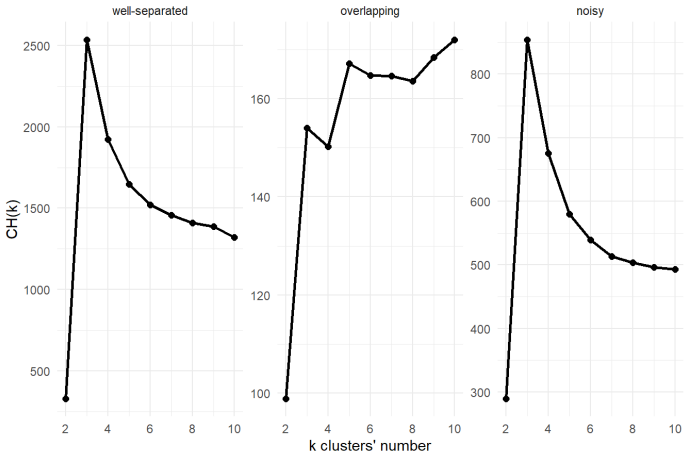


Fig. 5. Comparison of Calinski–Harabasz profiles for the well-separated, overlapping and noisy scenarios.

indicates that, when the clusters are compact and clearly separated, the ratio between between-cluster dispersion and within-cluster dispersion provides an accurate and unambiguous criterion for dendrogram cutting.

In the noisy scenario, the maximum of the CH index also occurs at $k = 3$, despite the presence of random observations. Although the absolute value of the index decreases with respect to the well-separated case, the maximum remains clearly identifiable. This suggests that the CH criterion is reasonably robust to moderate levels of noise when the main cluster structure remains sufficiently separated.

The overlapping scenario shows a substantially different behaviour. In this case, the CH index reaches its maximum at $k = 10$, overestimating the true number of clusters. Moreover, the CH curve is much flatter and contains several local maxima of similar magnitude. This indicates that, under overlapping conditions, the index may favour finer partitions that reduce within-cluster dispersion by subdividing ambiguous regions of the data. As a result, the selected partition may reflect local density fluctuations rather than the underlying generative structure.

These results show that the usefulness of the CH criterion should not be assessed only by the location of its maximum. The global morphology of the CH profile also provides relevant information about the reliability of the selected partition. Sharp and isolated maxima are associated with well-defined clustering structures, whereas flat profiles with several competing local maxima suggest ambiguous or overlapping structures.

6 Discussion

The experimental results show that the behaviour of the Calinski–Harabasz index across dendrogram α-cuts is strongly conditioned by the structure of the data. The index performs accurately when the clusters are clearly separated, remains stable under moderate noise contamination, and becomes less reliable when the underlying groups overlap.

This behaviour can be explained from the original definition of the Calinski–Harabasz index [1]. The criterion evaluates the ratio between between-cluster dispersion and within-cluster dispersion. In well-separated datasets, these two objectives are aligned with the true cluster structure: the partition with three groups produces compact clusters and a large between-cluster separation. This explains the sharp maximum observed at the true number of clusters.

In contrast, when clusters overlap, the true generative groups are no longer clearly separated in the feature space. Under these conditions, subdividing the overlapping regions may substantially reduce the within-cluster sum of squares without producing a proportionally large penalty in between-cluster separation. Consequently, the CH index may favour partitions with a larger number of clusters. In this sense, the criterion does not necessarily recover the generative number of groups, but rather the partition that optimizes a geometric compactness-separation trade-off.

This observation is particularly relevant for automatic dendrogram cutting. Selecting the partition associated with the maximum CH value can be effective when the CH curve contains a clear and isolated peak. However, when the profile is flat or contains several local maxima of similar magnitude, the selected cut should be interpreted with caution. In such cases, the maximum CH value may not represent a stable or semantically meaningful clustering solution.

The results also suggest that the complete CH profile contains information that is lost when only the maximum value is considered. A sharp maximum indicates that one partition is clearly preferred over neighbouring cuts, whereas a flat profile indicates that several alternative partitions provide similar values of the compactness-separation trade-off. This interpretation is related to broader notions of cluster stability and reliability discussed in the clustering literature [16]. Therefore, the shape of the CH curve can be interpreted as an additional diagnostic element for assessing the reliability of automatic dendrogram cuts.

The analysis of dendrogram height variations provides a complementary perspective. Large jumps in dendrogram height indicate important agglomeration events in the hierarchy and are commonly associated with meaningful cuts. When the maximum CH value and the largest height variation point to similar partitions, the evidence supporting the selected cut becomes stronger. Conversely, discrepancies between both criteria may indicate ambiguous hierarchical structure, overlap between groups or the presence of local substructures.

These findings are consistent with previous work by Karna and Gibert [15], who showed that the behaviour of the CH criterion across consecutive dendrogram partitions provides useful information for automatic tree cutting. The present study extends this perspective by showing that not only the value of the

index, but also the morphology of its complete profile, is relevant for interpreting the quality and reliability of the selected partition.

Despite these promising results, the study has some limitations. First, the experiments were conducted on synthetic two-dimensional numerical datasets in order to allow controlled interpretation and graphical inspection. Second, only Ward's linkage criterion with Euclidean distance was considered. This choice is appropriate because Ward's method is directly related to the minimization of within-cluster variance, which is also central to the CH index. However, further experiments are required to evaluate whether the observed behaviours remain valid under other linkage criteria, distance measures, dimensionalities and real-world datasets. Nevertheless, hierarchical clustering methodologies similar to those analysed in this work have already demonstrated their usefulness in real-world applications such as industrial condition monitoring and health assessment [12].

Overall, the results support the use of the Calinski–Harabasz index as an automatic dendrogram cutting criterion, but they also show that its maximum should not be interpreted in isolation. The complete CH profile should be examined jointly with dendrogram height variations to distinguish between reliable, well-defined solutions and ambiguous situations in which several competing partitions are plausible.

7 Conclusions

This paper presented an experimental study of the behaviour of the Calinski–Harabasz index across the sequence of α-cuts generated by hierarchical clustering dendrograms.

The results show that the CH index correctly identifies the true number of clusters when the underlying structure is clearly separated. The criterion also remains robust under moderate noise contamination, although the absolute value of the index decreases. In contrast, overlapping structures generate flatter CH profiles, several competing local maxima and possible overestimation of the number of clusters.

The main conclusion of this study is that the diagnostic value of the CH criterion extends beyond the location of its maximum. The global shape of the CH profile provides useful information about the reliability of the selected partition. Sharp and dominant maxima are associated with well-defined clustering structures, whereas flatter profiles suggest ambiguity, overlap or the presence of competing hierarchical solutions.

These findings support the use of the CH index as a practical criterion for automatic dendrogram cutting, especially when combined with an analysis of the complete index profile and dendrogram height variations. Future work will extend this analysis to heterogeneous and mixed-type datasets, additional linkage criteria, higher-dimensional scenarios and real-world clustering problems.

Acknowledgments. This research has been partially financed by the SGR2021-01532 funds from AGAUR, the predoctoral fellowship 2023-FISDU-00366 co-financed by the European Union.



References

1. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974)
2. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (1963)
3. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
4. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **1**(2), 224–227 (1979)
5. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics **4**(1), 95–104 (1974)
6. Milligan, G.W., Cooper, M.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika **50**(2), 159–179 (1985)
7. Gibert, K., Nonell, R., Velarde, J., Colillas, M.: Knowledge discovery with clustering: impact of metrics and reporting phase by using KLASS. Neural Network World **15**(4), 319–340 (2005)
8. Gibert, K., Sevilla-Villanueva, B., Sanchez-Marre, M.: The role of significance tests in consistent interpretation of nested partitions. Journal of Computational and Applied Mathematics **292**, 623–633 (2016)
9. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967)
10. Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: an overview. WIREs Data Mining and Knowledge Discovery **2**(1), 86–97 (2012)
11. Sevilla-Villanueva, B., Gibert, K., Sanchez-Marre, M.: Using CVI for understanding class topology in unsupervised scenarios. Conference of the Spanish Association for Artificial Intelligence, 135–149 (2016)
12. Gibert, K., Marti-Puig, P., Cusido, J., Sole-Casals, J.: Identifying health status of wind turbines using self-organizing maps and interpretation-oriented post-processing tools. Energies **11**(4), 723 (2018)
13. Gibert, K., Sanchez-Marre, M., Codina, V.: Choosing the right data mining technique: classification of methods and intelligent recommendation. International Congress on Environmental Modelling and Software (2010)
14. Karna, A., Gibert, K.: Using hierarchical clustering to understand behavior of 3D printer sensors. International Workshop on Self-Organizing Maps, 150–159 (2019)
15. Karna, A., Gibert, K.: Automatic identification of the number of clusters in hierarchical clustering. Neural Computing and Applications **34**, 119–134 (2022)
16. Hennig, C.: Cluster-wise assessment of cluster stability. Computational Statistics and Data Analysis **52**(1), 258–271 (2007)
17. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems **17**(2–3), 107–145 (2001)