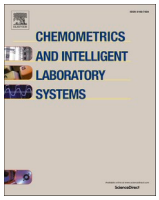
Chemometrics and Intelligent Laboratory Systems 251 (2024) 105117

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Extended multivariate comparison of 68 cluster validity indices. A review

Roberto Todeschini, Davide Ballabio, Veronica Termopoli, Viviana Consonni *

Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.zza della Scienza 1, 20126, Milano, Italy

ARTICLE INFO

Keywords:
Cluster analysis
Cluster validity indices
k-means algorithm
Minimum Spanning Tree
Principal Component Analysis
Machine learning
Data mining
Artificial intelligence

ABSTRACT

Clustering is an unsupervised machine learning methodology widely used in several sciences to find groups of similar patterns in complex data. The results generated by clustering algorithms generally depend on user-defined input parameters such as the number of expected clusters, which can have a great impact on the homogeneity of the identified clusters.

Clustering validity indices (CVIs) are an effective method for determining the optimal number of clusters that best fit the natural partition of a dataset. They do not require any underlying assumption nor a priori knowledge about the true dataset structure. Since 1965, many cluster validity indices have been proposed in the literature and used in several different applications.

In this paper, the performance of 68 cluster validity indices was evaluated on 21 real-life research and simulated datasets. CVIs were compared on the same partition for each dataset, which was searched for by the *k*-means clustering algorithm. Multivariate chemometric methods were applied to disclose mutual relationships among the indices and to select those that are more effective in terms of accuracy and reliability.

1. Introduction

In most of data analysis applications, there is no external criterion or knowledge to define some meaningful categories of the objects to study. They are characterized by a set of measurements and on the basis of the similarity between data points one can only attempt to characterize the structure of the dataset as best as possible. Cluster analysis helps to discover the 'natural' grouping of a set of data by unsupervised learning methods. Clustering approaches are a kind of exploratory data analysis, widely used to find groups of similar patterns in several research fields, such as science, medicine, engineering and social sciences for different purposes such as data categorization, information retrieval, web and text mining, image analysis, object recognition. A clustering algorithm generates a partition of the objects, which are represented as points in a *p*-dimensional space, into a number *K* of groups (i.e., commonly referred to as the parameter *K*), generally searching for homogeneity within the clusters and heterogeneity among different clusters [1].

Clustering algorithms can be divided into crisp and fuzzy methods. Crisp clustering (or hard clustering) assigns each data point to one and only one of the clusters, while fuzzy clustering allows each data point to belong to more than one cluster with a different membership degree. Then, in crisp clustering well defined boundaries are assumed among the clusters, while in fuzzy clustering they reflect the likely overlap of the groups, which can be often encountered in real datasets due to data uncertainty.

In general, there is no a unique "natural" partition of the objects, hence a reasonable purpose is to achieve a reliable solution that reveals some meaningful data patterns. Thus, the main questions to be answered are "how many clusters are there in the dataset?" and "how to find the 'natural' number of clusters?". Since most of the clustering algorithms require this information to be known in advance, the common approach is to run the algorithm several times with different values of the parameter *K*, compare the obtained partitions and choose the partition that best fits the data structure. This process, which implies a quantitative evaluation of the clustering results, is known under the general term of cluster validation [2]. There is a distinction between internal and external validation depending on the kind of information available in the validation process. External validation methods evaluate the clustering results by using, if available, the correct partition of the objects (i.e., the "true" data classes) and the indices for partition comparison, which are commonly referred to as partition similarity measures (e.g., Rand index, adjusted Rand index, Jaccard index, Fowlkes–Mallows index). Internal validation methods just examine the obtained data partition accounting for some specific features of the

* Corresponding author.
E-mail address: viviana.consonni@unimib.it (V. Consonni).

https://doi.org/10.1016/j.chemolab.2024.105117
Received 13 September 2023; Received in revised form 20 February 2024; Accepted 21 March 2024
Available online 25 March 2024

clusters (e.g., compactness, separation, density, overlap), which are used to calculate the internal cluster validity indices (CVIs). Among the cluster features considered in these indices, there are the connectedness and the external isolation. The former measures the cluster internal cohesion or density and relates to the number of connections within the cluster, the minimum similarity or the average similarity between the objects in the same cluster. The external isolation measures how well-separated are the objects within a cluster from the objects in other clusters and can be calculated as the distance between the two nearest objects that belong to different clusters or as the average distance between all the objects belonging to different clusters or else as the smallest distance between the cluster centroids.

This study focuses on the internal validation since in most of data mining applications the underlying structure of the data is unknown. The importance of this research field is also evidenced by the large number of software packages for cluster validity indices calculation, which have been developed in the last 15 years: clValid [3], ccCrit [4], cclust [5], clusterSim [6], NbClust [7], clv [8].

Previous comparisons among the most common cluster validity indices can be found in the papers of Milligan [9,10], Halkidi et al. [11,12], Bandyopadhyay et al. [13,14], Pakhira et al. [15], Kim et al. [16], Tang et al. [17], Wu et al. [18], Saitta et al. [19], Zhang et al. [20], Saha et al. [21], Sengupta et al. [22], Arbelaitz et al. [23], Brito da Silva et al. [24], Wiroonsri [25]. In the majority of the cases, only a small number of cluster validity indices have been compared and/or on a few number of datasets and, in none of these studies, a multivariate comparison has been carried out.

The purpose of our study is to survey most of the validity indices (CVIs) that have been proposed so far for crisp clustering and compare their performance on a large number of different real-life and simulated datasets by a multivariate chemometric perspective. The remainder of the paper is structured as follows. The first section deals with the theoretical fundamentals and, in particular, it describes the cluster validity indices through their formal mathematical definitions and selection rules. In the second section, the relevant features of the datasets are presented along with the methodology used to generate the set of partitions for the calculation of the CVIs. Section three provides the most relevant results of the index comparison and an application of the CVIs to a real-life dataset with complex data structure.

2. Materials and methods

2.1. Theoretical fundamentals of the cluster validity indices

2.1.1. Algebraic notation and formal definitions

We will use the following notation to describe the computational formulae of the cluster validity indices.

The cluster validity index (CVI) is a numerical valued function defined for all the partitions of the objects into a varying number K of clusters; the optimal value $\widehat{K}$ of the number of clusters provided by the validity index can be formally represented as a function of three elements: $\widehat{K}(\mathrm{CVI}) = f(P, A, R)$.

- The element P is the set $P \equiv \{P_2, P_3, \ldots, P_{K^{max}}\}$ of $K^{max} - 1$ different partitions of the objects into a variable number K of clusters, which is progressively increased from 2 to an upper user-defined value K^{max}. The partition with $K = 1$ corresponds to the entire set of objects and it is not considered in this study due to the impossibility to calculate most of the CVIs and since this case is mainly related to the general problem of clusterizability of a dataset, which is out of the scope of this work.
- The element A is the algorithm applied to the partition P_K to calculate the K-th value of the cluster validity index: $\mathrm{CVI}_K = A(P_K)$, $K = 2, \ldots, K^{max}$.
- The element R is the stopping rule or criterion adopted to automatically determine the optimal number of clusters $\widehat{K}$; it involves observing the behaviour of the internal measure of cluster validity (i.e., CVI) as the number K of clusters is increased from 2 to the maximum allowed value and selecting the appropriate value K for which the numerical value of the validity index is optimal.

n	number of objects
p	number of variables
K	number of clusters
n_k	number of objects in the k-th cluster
I_k	the set of the numerical identifiers of the objects belonging to the k-th cluster
$\mathbf{X}(n \times p)$	data matrix of the entire set of objects
$\mathbf{X}_k(n_k \times p)$	data matrix of the k-th cluster
$\mathbf{x}_i$	p-dimensional row vector of the i-th object
$\mathbf{b} = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{x}_i$	p-dimensional row vector of the variable means of the dataset (barycentre)
$\mathbf{c}_k$	p-dimensional row vector of the variable means of the k-th cluster (centroid)
$\bar{\mathbf{c}} = \frac{1}{K} \cdot \sum_{k=1}^{K} \mathbf{c}_k$	p-dimensional row vector of the variable means of the global centroid
$\|\mathbf{x}\|$	Euclidean norm of a vector $\mathbf{x}$

The rule R is the final and decisive step for selecting the best partition of the objects, that is, the partition that best fits the natural data patterns. In several cases, depending on the theoretical definition of the index, the optimal number of clusters is given by the minimum (or maximum) value of the CVI. However, this decision is not always simple due to the monotonic (ascending or descending) or degenerative trend of some CVIs. In these cases, it is necessary to modify the rule in such a way as trivial solutions and/or problems of human subjectivity are avoided. Then, the first maximum, or first minimum, is a better alternative than the max/min criterion to avoid potential not relevant absolute maxima or minima due to some degenerative behaviour of the index after the optimal K value. Moreover, the value K at which a marginal change in the index from one clustering level to the next is observed to flatten drastically may indicate that further division of the clusters is not required since no significant improvement of the internal validity measure is obtained. This rule has been called *max ratio* and is formally defined as the following:

$$\widehat{K} = arg_K \, max\left(\frac{\mathrm{CVI}_K - \mathrm{CVI}_{K-1}}{\mathrm{CVI}_{K+1} - \mathrm{CVI}_K}\right) \tag{1}$$

The extreme values of the cluster validity index, that is, CVI_1 and $\mathrm{CVI}_{K^{max}+1}$, which are necessary to calculate CVI_2 and $\mathrm{CVI}_{K^{max}}$, were estimated by a spline interpolation.

Validity indices aim to quantitatively evaluate the clustering results and are useful to select the best partition of the dataset and more specifically the optimal number K of clusters, which is commonly known as the parameter K. Validity indices were compared on the same partition of each dataset and, thus, the index sensitivity or stability to the different data partitions that can result from the clustering method iterations for a given K were not considered in this study to avoid an additional source of data variation. An optimal data partition should have well separated clusters whose members are very similar to each other. Then, cluster validity indices are based on two fundamental concepts [12]. The first one relies on the cluster homogeneity and is what is usually called *compactness* (or cohesion, tightness, connectedness), that is, the extent to which the members of each cluster are close to each other in the descriptor space. Common measures of compactness are the within-group dispersion, or variance, and the intra-cluster distance, which should be minimized. The second concept is *separation*, that is, the extent to which the clusters are far apart in the descriptor space. There are some common approaches to measure the separation between two clusters: 1) the nearest neighbour distance, that is, the distance between the closest members of two different clusters (i.e., single linkage); 2) the farthest neighbour distance, that is, the distance between the most distant members (i.e., complete linkage); 3) the distance between the centroids of the clusters (i.e., average linkage).

2.1.2. Indices based on dispersion measures

Most of the well-known cluster validity indices (Tables 1 and 2) are designed to account for some basic dispersion measures that can be calculated from the matrices described below.

The total scatter matrix **T** measures the dispersion around the barycentre **b** of the data matrix **X**:

$$\begin{aligned}&\mathbf{T}(p\times p)=(\mathbf{X}-\mathbf{b})^{\mathrm{T}}\cdot(\mathbf{X}-\mathbf{b})\\&\text{where }[\mathbf{T}]_{jq}=\sum_{i=1}^{n}\left(x_{ij}-b_j\right)\cdot\left(x_{iq}-b_q\right)\end{aligned}\tag{2}$$

The trace of **T** is the sum of all the squared Euclidean distances from the data barycentre and it is known as the total sum of squares (*TSS*):

$$tr(\mathbf{T})=\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij}-b_j\right)^2=\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{b}\|^2=TSS\tag{3}$$

The within-group scatter matrix $\mathbf{WG}_k$ accounts for the dispersion of the k-th cluster objects around their centroid $\mathbf{c}_k$:

$$\begin{aligned}&\mathbf{WG}_k(p\times p)=(\mathbf{X}_k-\mathbf{c}_k)^{\mathrm{T}}\cdot(\mathbf{X}_k-\mathbf{c}_k)\\&\text{where }\ [\mathbf{WG}]_{jq}=\sum_{i\in I_k}\left(x_{ij}-c_{kj}\right)\cdot\left(x_{iq}-c_{kq}\right)\end{aligned}\tag{4}$$

The trace of the within-group scatter matrix, sometimes denoted as *WGSS*, is the sum of the squared Euclidean distances from the cluster centroid:

$$tr(\mathbf{WG}_k)=\sum_{j=1}^{p}\sum_{i\in I_k}\left(x_{ij}-c_{kj}\right)^2=\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|^2=WGSS\tag{5}$$

The division of *WGSS* by n_k (i.e., the number of objects in the cluster) provides the average squared intra-cluster distance or cluster variance.

The pooled within-group scatter matrix **WG** is the within-cluster sum of squares and cross products matrix; it can be calculated by adding the individual cluster scatter matrices $\mathbf{WG}_k$ over all the clusters as:

$$\mathbf{WG}(p\times p)=\sum_{k=1}^{K}\mathbf{WG}_k\tag{6}$$

The matrix **WG** encodes the pooled amount of data variation that is present in each cluster and thus, it provides information on the degree of similarity or homogeneity of the objects in each cluster. More specifically, the trace of **WG** gives the within-group sum of squares (*WSS*), which measures the pooled within-cluster cohesion, that is, the sum of the squared distances of the objects from the centroid of the cluster they belong to:

$$tr(\mathbf{WG})=\sum_{k=1}^{K}\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|^2=WSS\tag{7}$$

Finally, the between-group scatter matrix **BG** is the between-cluster sum of squares and cross products matrix; this matrix gives information about the extent to which clusters are different from each other and can be simply obtained as the difference between the total and the within-group scatter matrices as:

$$\begin{aligned}&\mathbf{BG}(p\times p)=\mathbf{T}-\mathbf{WG}\\&\text{where }[\mathbf{BG}]_{jq}=\sum_{k=1}^{K}n_k\cdot\left(c_{kj}-b_j\right)\cdot\left(c_{kq}-b_q\right)\end{aligned}\tag{8}$$

$$tr(\mathbf{BG})=\sum_{j=1}^{p}\sum_{k=1}^{K}n_k\cdot\left(c_{kj}-b_j\right)^2=\sum_{k=1}^{K}n_k\cdot\|\mathbf{c}_k-\mathbf{b}\|^2=BSS\tag{9}$$

where *BSS* is the between-group sum of squares.

The first two cluster validity indices of Table 1 date back to 1965. The index Trace_W (**1**) is the total cluster dispersion measured by the within-group sum of squares (*WSS*) and over the years it has been one of the most common validity indices in clustering applications [26]; Ball-Hall BH index (**2**) is its average counterpart [27]. The Banfield-Raftery BR index (**3**) was proposed in 1993 [28] as an alternative to the sum of squares criterion (*WSS*). While *WSS* is likely to perform well when all the clusters have the same dispersion [29], the index BR, which is based on the sum of the average squared distances from the cluster centroids (i.e., cluster variances), tends to be more appropriate when the clusters are hyperspherical but of different sizes; the size of a cluster is intended as the volume occupied by the cluster in the multivariate space rather than the number of objects it contains.

Calinski-Harabasz CH index (**4**) is a classical cluster validity index proposed in 1974 [30] as the ratio of the between-cluster to the within-cluster variance following the rationale of a pseudo ANOVA F test. In 1975, Hartigan [31] proposed a logarithmic scale-based variant of CH, that is, the LSSR index (**6**), which is defined as the logarithmic of the ratio of the sum of the between-cluster squared distances (*BSS*) to the sum of the squared within-cluster distances (*WSS*). Ratkowsky-Lance RL index (**7**), proposed in 1979 [32], is based on the ratio of the sum of the squared between-cluster distances to the sum of the squared distances in the entire dataset, but considering the average of the ratios calculated for each variable x of the dataset. Some years later, in 1996, Sharma [33] defined a similar index RS (**8**) as the ratio of *BSS* to *TSS*, which ranges from 0 (i.e., no difference among groups) to 1 (i.e., maximum difference among groups). This index measures the extent to which clusters are different from each other or, alternatively, the extent to which they are homogeneous, since the larger the *BSS* the smaller the *WSS* and vice versa.

A variant of Calinski-Harabasz index (**4**), which was proposed by Zhu et al. in 2019 [34] to more efficiently process datasets with large overlap among clusters, is the index WCH (**5**). This index was designed to account for three features: like the index CH, it measures the cluster compactness by the within-cluster variance and the inter-cluster separation by the between-cluster variance; in addition, a correction factor accounts for the inter-cluster overlap of the dataset.

The Davies-Bouldin index, proposed in 1979 [35] and here denoted as DB1 (**9**), is defined as the average of the overlap measure of each cluster with other clusters, which relates the within-cluster dispersion to the inter-cluster separation. Each cluster is compared to all the other clusters and associated the maximum ratio of the sum of the radii of the two considered clusters to the distance between their centroids. The radius accounts for the cluster size and is calculated as the average distance of the objects in the same cluster from their centroid. A variant of this index, denoted as DB2 (**10**), was proposed in 2005 [16] as the average of the sum over all the clusters of the ratio of the largest sum of two cluster radii to the smallest distance between two cluster centroids.

The Pakhira–Bandyopadhyay–Maulik PBM index (**11**), also called the *I*-index, is based on three factors [15]: 1) the first factor accounts for the comparison between the total scatter of the dataset, which is as considering all the objects belonging to one single cluster, and the total within-cluster dispersion after the objects are partitioned into a number of clusters; the ratio of these two quantities tends to increase with the increasing of the number of clusters; 2) the second factor is the maximum distance between cluster centroids, which remains constant after a certain value of K; 3) the third factor is the inverse of the number K of clusters, which was introduced to compensate for the growth of the dispersion ratio with further data partitioning.

The Fukuyama-Sugeno FS index (**16**) was proposed in 1989 as a new validity index for the fuzzy c-means method [36]. It is defined as the difference between two terms; the first term is a compactness measure (i.e., *WSS*) and the second term is the degree of separation between each cluster and the mean of the cluster centroids ($\bar{\mathbf{c}}$).

Table 1
Cluster validity indices based on dispersion measures. The second to last column reports the rule to search for the optimal index value.

ID	Index (pub. year)	Formula	Rule	Ref.
1	Trace_W (1965)	$\mathrm{trW} \equiv WSS = \sum_{k=1}^{K}\sum_{i\in I_k}\|\mathbf{x}_i - \mathbf{c}_k\|^2$	max ratio	[26]
2	Ball-Hall (1965)	$\mathrm{BH} = \frac{1}{K}\cdot\sum_{k=1}^{K}\sum_{i\in I_k}\|\mathbf{x}_i - \mathbf{c}_k\|^2 / n_k$	max ratio	[27]
3	Banfield-Raftery (1993)	$\mathrm{BR} = \sum_{k=1}^{K} n_k \cdot log\left(\sum_{i\in I_k}\|\mathbf{x}_i - \mathbf{c}_k\|^2 / n_k\right)$	max ratio	[28]
4	Calinski-Harabasz (1974)	$\mathrm{CH} = \frac{BSS/(K-1)}{WSS/(n-K)}$	first max	[30]
5	WCH (2019)	$\mathrm{WCH} = \frac{BSS/(K-1)}{WSS/(n-K) + \sum_{k=1}^{K-1}\sum_{k=k+1}^{K} f_{kk'}/n}$ $f_{kk'} = \begin{cases} 1 & if \ \ \|\mathbf{c}_k - \mathbf{c}_{k'}\|^2 < \sum_{s\in I_k}\|\mathbf{x}_s - \mathbf{c}_k\|^2/n_k + \sum_{t\in I_{k'}}\|\mathbf{x}_t - \mathbf{c}_{k'}\|^2/n_{k'} \\ 0 & otherwise \end{cases}$	max	[34]
6	Hartigan (1975)	$\mathrm{LSSR} = log\left(\frac{BSS}{WSS}\right)$	max ratio	[31]
7	Ratkowsky-Lance (1978)	$\mathrm{RL} = \sqrt{\frac{1}{K\cdot p}\cdot\sum_{j=1}^{p}\frac{\sum_{k=1}^{K} n_k\cdot(c_{kj} - b_j)^2}{\sum_{i=1}^{n}(x_{ij} - b_j)^2}}$	max ratio	[32]
8	R-Squared (1996)	$\mathrm{RS} = \frac{BSS}{TSS} = \frac{BSS}{WSS + BSS}$	max ratio	[33]
9	Davies-Bouldin (1979)	$\mathrm{DB1} = \frac{1}{K}\cdot\sum_{k=1}^{K} max_{k'\neq k}\left(\frac{\sum_{s\in I_k}\|\mathbf{x}_s - \mathbf{c}_k\|/n_k + \sum_{t\in I_{k'}}\|\mathbf{x}_t - \mathbf{c}_{k'}\|/n_{k'}}{\|\mathbf{c}_k - \mathbf{c}_{k'}\|}\right)$	min	[35]
10	Davies-Bouldin* (2005)	$\mathrm{DB2} = \frac{1}{K}\cdot\sum_{k=1}^{K}\frac{max_{k'\neq k}\left(\sum_{s\in I_k}\|\mathbf{x}_s - \mathbf{c}_k\|/n_k + \sum_{t\in I_{k'}}\|\mathbf{x}_t - \mathbf{c}_{k'}\|/n_{k'}\right)}{min_{k'\neq k}\|\mathbf{c}_k - \mathbf{c}_{k'}\|}$	min	[16]
11	Pakhira-Bandyopadhyay-Maulik (2001)	$\mathrm{PBM} = \left(\frac{1}{K}\cdot\frac{\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{b}\|}{\sum_{k=1}^{K}\sum_{i\in I_k}\|\mathbf{x}_i - \mathbf{c}_k\|}\cdot max_{k\neq k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|\right)^2$	max	[15]
12	Žalik SV (2011)	$\mathrm{SV} = \frac{\sum_{k=1}^{K} min_{k'\neq k}\|\mathbf{c}_{k'} - \mathbf{c}_k\|}{\sum_{k=1}^{K} max_{i\in I_k}\|\mathbf{x}_i - \mathbf{c}_k\|}$	first max	[41]
13	Wemmert-Gancarski (2013)	$\mathrm{WG1} = \frac{1}{n}\cdot\sum_{k=1}^{K} max\left(0, \ \ n_k - \sum_{i\in I_k}\frac{\|\mathbf{x}_i - \mathbf{c}_k\|}{min_{k'\neq k}\|\mathbf{x}_i - \mathbf{c}_{k'}\|}\right)$	max	[4]
14	Wemmert-Gancarski* (2023)	$\mathrm{WG2} = \frac{1}{K}\cdot\sum_{k=1}^{K}\frac{1}{n_k}\cdot\sum_{i\in I_k}\left(1 - \frac{\|\mathbf{x}_i - \mathbf{c}_k\|}{min_{k'\neq k}\|\mathbf{x}_i - \mathbf{c}_{k'}\|}\right)$	max	This work
15	Score Function (2007)	$\mathrm{SF} = 1 - \left\{exp\left[exp\left(\frac{1}{n\cdot K}\cdot\sum_{k=1}^{K} n_k\cdot\|\mathbf{c}_k - \bar{\mathbf{c}}\| - \sum_{k=1}^{K}\frac{1}{n_k}\cdot\sum_{i\in I_k}\|\mathbf{x}_i - \mathbf{c}_k\|\right)\right]\right\}^{-1}$	max	[19]
16	Fukuyama-Sugeno (1989)	$\mathrm{FS} = \sum_{k=1}^{K}\sum_{i\in I_k}\|\mathbf{x}_i - \mathbf{c}_k\|^2 - \sum_{k=1}^{K}\|\mathbf{c}_k - \bar{\mathbf{c}}\|^2$	max ratio	[36]
17	Xie-Beni (1991) – Ray-Turi (1999)	$\mathrm{XB1} \equiv \mathrm{RT} = \frac{WSS/n}{min_{k\neq k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2}$	min	[37,38]

(*continued on next page*)

Table 1 (*continued*)

ID	Index (pub. year)	Formula	Rule	Ref.
18	Xie-Beni* (2005)	$XB2 = \dfrac{max_k\left(\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|^2/n_k\right)}{min_{k\neq k'}\|\mathbf{c}_k-\mathbf{c}_{k'}\|^2}$	min	[16]
19	Kwon (1998)	$Kw = \dfrac{WSS+\sum_{k=1}^{K}\|\mathbf{c}_k-\mathbf{b}\|^2/K}{min_{k\neq k'}\|\mathbf{c}_k-\mathbf{c}_{k'}\|^2}$	min	[39]
20	Tang (2005)	$Tn = \dfrac{WSS+\frac{2}{K\cdot(K-1)}\cdot\sum_{k=1}^{K}\sum_{k'=k+1}^{K-1}\|\mathbf{c}_k-\mathbf{c}_{k'}\|^2}{min_{k\neq k'}\|\mathbf{c}_k-\mathbf{c}_{k'}\|^2+1/K}$	min	[17]
21	Partition Separation (2001)	$PS = \sum_{k=1}^{K}\left(\dfrac{n_k}{max_{k'}(n_{k'})}-exp\left[-\dfrac{min_{k'\neq k}\|\mathbf{c}_k-\mathbf{c}_{k'}\|^2}{\frac{1}{K}\cdot\sum_{k=1}^{K}\|\mathbf{c}_k-\bar{\mathbf{c}}\|^2}\right]\right)$	max	[40]
22	Rezaee-Lelieveldt-Reiber (1998)	$SD1 = \alpha\cdot Scat + Dis \qquad \alpha = Dis(K^{max})$ $Scat = \dfrac{1}{K}\cdot\sum_{k=1}^{K}\dfrac{\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|^2/n_k}{\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{b}\|^2/n}$ $Dis = \dfrac{max_{k\neq k'}\|\mathbf{c}_k-\mathbf{c}_{k'}\|}{min_{k\neq k'}\|\mathbf{c}_k-\mathbf{c}_{k'}\|}\cdot\sum_{k=1}^{K}\left(\sum_{k=1;k'\neq k}^{K}\|\mathbf{c}_k-\mathbf{c}_{k'}\|\right)^{-1}$	min	[43]
23	Kim-Ramakrishna SD* (2005)	$SD2 = \alpha\cdot Scat^* + Dis \qquad \alpha = Dis(K^{max})$ $Scat^* = max_k\left(\dfrac{\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|^2/n_k}{\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{b}\|^2/n}\right)$	min	[16]
24	Halkidi (2001)	$SDbw = Scat + Dis^*$ $Dis^* = \dfrac{1}{K\cdot(K-1)}\cdot\sum_{k=1}^{K}\sum_{k=1;k'\neq k}^{K}\dfrac{N(\bar{\mathbf{c}}_{kk'})}{max(N(\mathbf{c}_k),N(\mathbf{c}_{k'}))} \qquad \bar{\mathbf{c}}_{kk'}=\dfrac{\mathbf{c}_k+\mathbf{c}_{k'}}{2}$ $N(\mathbf{c}_k)=\sum_{i\in I_k}f_i(\mathbf{c}_k) \qquad N(\bar{\mathbf{c}}_{kk'})=\sum_{i\in I_k\cup I_{k'}}f_i(\bar{\mathbf{c}}_{kk'})$ $f_i(\mathbf{c}_k)=\begin{cases}1 & \text{if } \|\mathbf{x}_i-\mathbf{c}_k\|\leq\sigma\\ 0 & \text{if } \|\mathbf{x}_i-\mathbf{c}_k\|>\sigma\end{cases} \qquad \sigma=\frac{1}{K}\cdot\sqrt{\sum_{k=1}^{K}\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|^2/n_k}$	first min	[12]
25	Kim-Park v_{sv} (2001)	$Vsv1 = v_{uN}+v_{oN}$ $v_u=\dfrac{1}{K}\cdot\sum_{k=1}^{K}\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|/n_k \qquad v_o=\dfrac{K}{min_{k\neq k'}\|\mathbf{c}_k-\mathbf{c}_{k'}\|}$	first min	[16,44]
26	Kim-Ramakrishna v_{sv}* (2005)	$Vsv2 = v^*_{uN}+v_{oN}$ $v^*_u = max_k\left(\sum_{i\in I_k}\|\mathbf{x}_i-\mathbf{c}_k\|/n_k\right)$	first min	[16]

The Xie-Beni index, here denoted as XB1 (**17**), is another ratio-type validity index originally proposed in 1991 [37] for fuzzy clustering, which uses the global average squared distance of the objects from their cluster centroid as the measure of cluster cohesion in the numerator and the minimum squared distance between pairs of clusters as the inter-cluster separation measure in the denominator. The Ray-Turi index RT has the same definition as the Xie-Beni index but it does not consider the fuzzy membership of objects [38]. In the case of crisp clustering, the two indices coincide and for this reason, only one of them (i.e., XB1) has been considered in this comparative study.

In 1998, Kwon [39] proposed the index Kw (**19**) as a modification of the Xie-Beni index (**17**) to overcome its monotonic decreasing tendency as the number of clusters becomes very large and near the number of objects, by introducing an *ad hoc* penalty function defined as the average squared distance of the clusters from the barycentre. Following the same idea as Kwon, the Tang index (**20**) is another variant of the Xie-Beni index with the introduction of two penalty functions [17]: 1) a first penalty function, defined as the average squared distance between the cluster centroids, is added to the term in the numerator to adjust the decreasing tendency and 2) a second penalty function, defined as the reciprocal of the number of clusters, is added to the term in the denominator with the aim to strengthen the numerical stability as the membership weighting exponent increases when the index is used in the fuzzy version. The variant of the Xie-Beni index proposed by Kim and

Table 2
Cluster validity indices based on dispersion matrix algebraic operators. Notation *tr* indicates the matrix trace and $|\mathbf{X}|$ refers to the matrix determinant. The second to last column indicates the rule to search for the optimal index value.

ID	Index (pub. year)	Formula	Rule	Ref.
27	Friedman-Rubin 1 (1967)	$\text{trWB} = tr\left(\frac{\mathbf{BG}}{\mathbf{WG}}\right)$	max ratio	[45]
28	Friedman-Rubin 2 (1967)	$\text{DR} = \frac{\lvert\mathbf{T}\rvert}{\lvert\mathbf{WG}\rvert}$	max ratio	[45]
29	Scott-Symons 1 (1971)	$\text{SS} = \sum_{k=1}^{K} n_k \cdot log\left\lvert\frac{\mathbf{WG}_k}{n_k}\right\rvert$	max ratio	[47]
30	Scott-Symons 2 (1971)	$\text{LDR} = n \cdot log\left(\frac{\lvert\mathbf{T}\rvert}{\lvert\mathbf{WG}\rvert}\right)$	max ratio	[47]
31	Marriot (1975)	$\text{KDW} = K^2 \cdot \lvert\mathbf{WG}\rvert$	first min	[46]
32	Fuzzy HyperVolume (1989)	$\text{FHV} = \sum_{k=1}^{K} \left\lvert\frac{\mathbf{WG}_k}{n_k}\right\rvert^{1/2}$	first min	[48]
33	Negentropy Increment (2010)	$\text{NI} = \frac{1}{2} \cdot \sum_{k=1}^{K} \frac{n_k}{n} \cdot log\left\lvert\frac{\mathbf{WG}_k}{n_k - 1}\right\rvert - \frac{1}{2} \cdot log\left\lvert\frac{\mathbf{T}}{n-1}\right\rvert - \sum_{k=1}^{K} \frac{n_k}{n} \cdot log\left(\frac{n_k}{n}\right)$	first min	[1]

Ramakrishna [16], here denoted as XB2 (**18**), replaces the global average measure of cluster compactness at the numerator with the maximum cluster variance since averaging generally tends to hide the effect due to unnecessary merging of clusters.

The partition separation PS index (**21**) was proposed in 2001 [40] for fuzzy clustering and here proposed in its crisp version. For each cluster, it combines a measure of cluster size, which is the proportion of cluster objects with respect to the cluster with the largest number of objects, and an exponential normalized separation measure defined in terms of the minimum squared distance from the other cluster centroids. In the original version for fuzzy clustering, the first factor was a normalized partition coefficient accounting for the cluster membership values of all the objects.

The index SV (**12**) was proposed with the aim to efficiently validate data partitions characterized by the presence of clusters that widely differ in size and density [41]. Like Dunn's index GDI11 (**42**) it is calculated as the ratio of an inter-cluster separation measure to a compactness measure; the compactness is evaluated considering the average distance of only the ten percent of the objects that are the farthest objects from the cluster centroids, while the separation measure is the sum of the smallest pairwise distances between cluster centroids.

The Wemmert-Gancarski WG1 index (**13**) has been described by Desgraupes [4]. It is based on a cluster score that accounts for the number of objects that are closer to their cluster centroid than to the centroids of other clusters. A variant of this index, denoted by WG2 (**14**), has been proposed in this study in analogy with the underlying idea of the Silhouette index [42], which defines a cluster membership score for each object. The membership score is calculated by comparing the distance of each object from the centroid of its cluster and its minimum distance from the other cluster centroids.

The Score Function SF (**15**) was proposed in 2007 by Saitta et al. [19] as a bounded validity index able to measure the proximity of the calculated partition to the ideal case of highly compacted and well-isolated clusters, for which the SF index reaches its maximum value of 1. In addition, unlike some other CVIs, it can handle the particular case of one cluster partition. This index is an exponential function of the difference between the separation measure, which is the cluster size-weighted average distance of the clusters from the overall cluster centroid, and the compactness measure, which is the sum over all the clusters of the average within-cluster distances.

The index SD1 (**22**) was proposed in 1998 by Rezaee, Lelieveldt and Reiber [43] in the framework of fuzzy c-means clustering and originally denoted by V_{CWB}, where the subscript CWB means Compose Within and Between scattering; it is a summation-type index that combines in an additive way the measures of cluster compactness (*Scat*) and separation (*Dis*). It was later adapted to crisp clustering and called SD by Halkidi, Vazirgiannis and Batistakis [12]. The first term *Scat* of this index represents the average of normalized variances within the clusters. The second term *Dis* indicates the total separation between the clusters and, generally, it is sensitive to both the number of clusters and the geometry of the cluster centres. Since the two terms vary in a different range of values, the weighting factor α, which is the term *Dis* at the maximum allowed number of clusters K^{max}, has been introduced in order to counterbalance both terms in a proper way. In ideal conditions, this index assumes that the measure of cluster compactness has a steep increase when the number K of clusters decreases from the optimal K^* value to $K^* - 1$ due to unnecessary cluster merging, while the inter-cluster separation decreases sharply when K decreases from $K^* + 1$ to K^*. Hence, the summation of these two terms has a minimum at K^* [16]. Based on the same design principles, variants of the index SD were later proposed using different measures of cluster compactness and separation [44]. SDbw (**24**) replaced the total separation with the density of the objects in the middle of two clusters and omitted the weighting factor [11]. The index Vsv1 (**25**) uses the average of the cluster mean absolute deviations as the first term v_u and the minimum inter-cluster distance as the second term v_o; v_{uN} and v_{oN} are the min–max normalized versions of v_u and v_o, respectively [16,44]. The index Vsv2 (**26**) is a variant of Vsv1, in which the first term representing the cluster compactness is calculated as the maximum cluster mean absolute deviation [16].

In analogy with trace_W (**1**), the index trWB (**27**), also called Hotelling's Trace and reported in Table 2, was introduced by Friedman and Rubin in 1967 [45] as the trace of the between-group to the within-group scatter matrix ratio. The other cluster validity indices collected in Table 2 are defined in terms of the determinant of some combination of the different scatter matrices. Among these, there is the determinant ratio DR (**28**) still proposed by Friedman and Rubin [45] as the ratio of the determinant of the total scatter matrix **T** to the determinant of the within-group scatter matrix **WG**. Since the determinant of the total scatter matrix is constant for a given dataset, they [45] also suggested as an alternative criterion the minimization of the determinant of the within-group scatter matrix $|\mathbf{WG}|$, which is almost the same criterion as the KDW index (**31**) later used by Marriot [46], who introduced the multiplying factor K^2 to improve the criterion ability to detect

the optimal number of clusters. Friedman and Rubin also suggested the use of the logarithmic function of the determinant of the within-group scatter matrix, which was revised by Scott and Symons in 1971 [47] to define the index LDR (**30**). In the same study [47], Scott and Symons defined a variant of this index, here denoted as SS (**29**), which accounts for the individual cluster scatter matrices instead of the pooled within-cluster scatter matrix. The index of Friedman and Rubin, based on the assumption that the within-group scatter matrix is the same for each cluster, tends to favour partitions with ellipsoidal clusters with the same orientation and size, whereas the index of Scott and Symons is able to account for clusters of different orientations, shapes, and sizes [28], but cannot be properly calculated in the case that clusters are not well represented.

Based on the same rationale of the Scott-Symons index (**30**), with the aim to account for the presence of large variability in cluster shapes, densities and number of objects in each cluster, the Fuzzy HyperVolume index (**32**) was proposed in 1989 by Gath and Geva [48]. This index adds over all the clusters the square root of the determinant of the cluster covariance matrix and was specifically designed for fuzzy clustering but here adopted in its crisp version allowing the membership function only to be 1 or 0 for objects belonging or not belonging to the cluster, respectively. In more recent years (2010), the Negentropy Increment (**33**) was introduced by Lago-Fernández and Corbacho assuming that a normally distributed cluster is optimal [1]. This index is based on the average normality of the clusters; the normality of a cluster is defined in terms of its negentropy, which measures the cluster deviation from normality and is calculated as the difference between the actual cluster entropy and the entropy of a normal distribution with the same covariance matrix.

2.1.3. Pairwise distance-based indices

Pairwise distance-based cluster validity indices are listed in Table 3. Their underlying idea is to describe the cluster compactness and the inter-cluster separation in terms of proximity (i.e., similarity/diversity) of the objects that belong to the same cluster and the objects belonging to different clusters, respectively. Similarity/diversity between objects is commonly measured by the Euclidean pairwise distance. Most of these distance-based indices do not account for the cluster shape, implicitly assuming that clusters are hyper-spheres.

Indices from ID 34 to 40, which mainly represent various types of correlation measures, are based on the following measures calculated from the pairwise distance matrix.

1) N^+ is the number of times a distance between two points not belonging to the same cluster is strictly greater than the distances between two points belonging to the same cluster (i.e., the number of concordant comparisons);
2) N^- is the number of times a distance between two points not belonging to the same cluster is strictly smaller than the distances between two points belonging to the same cluster (i.e., the number of discordant comparisons).
3) $n_T = n{\cdot}(n-1)/2$ is the total number of pairwise distances in the dataset, that is, the total number of distinct pairs of objects.
4) $n_W = \sum_{k=1}^{K} n_k{\cdot}(n_k-1)/2$ and $n_B = \sum_{k=1}^{K-1}\sum_{k=k+1}^{K} n_k{\cdot}n_{k'}$ are the number of within-cluster and the number of between-cluster pairwise distances, respectively.
5) $S_W = \sum_{k=1}^{K}\sum_{s<t\in I_k} \|\mathbf{x}_s - \mathbf{x}_t\|$ and $S_B = \sum_{k=1}^{K-1}\sum_{k=k+1}^{K}\sum_{s\in I_k, t\in I_{k'}} \|\mathbf{x}_s - \mathbf{x}_t\|$ are the sum of the distances between all the pairs of objects that belong to the same cluster and the sum of the distances between all the pairs of objects that belong to different clusters, respectively. S_{min} and S_{max} are the sum of the n_W smallest and n_W largest distances in the entire dataset, respectively [49]; they are used to calculate the C-Index (**38**).

Indices Tau (**34**), Gamma (**35**), G-plus (**36**) and G-minus (**37**) were first described in Ref. [50] as functions that measure the discrepancy between two dissimilarity matrices, that is, the original dissimilarity matrix and the cophenetic matrix. The elements of the cophenetic matrix are defined as the distance (or similarity) level at which two objects become members of the same cluster. These CVIs were proposed in the framework of hierarchical clustering methods. In the case of *k*-means partitioning, the cophenetic matrix can be still defined on the basis of some threshold value that allows to establish the group membership of the objects. Index Tau (**34**) is Kendall's rank correlation coefficient between the ranks, which are corrected for ties, assigned to the object pairs on the basis of their proximity, where similar object pairs are assigned the lower ranks, and the binary vector in which a value of 0 is assigned to a pair of objects that belong to the same cluster and a value of 1 to a pair of objects that belong to different clusters. The computational formula of Tau index is given in Ref. [9], where the term t in the denominator indicates the number of comparisons of two pairs of objects such that both pairs represent within cluster comparisons (i.e., within-cluster distances) or both pairs arc between cluster comparisons (i.e., between-cluster distances). Gamma index (**35**) is an adaptation of Goodman and Kruskal's Gamma correlation index [51] to be used for clustering applications [10]; this is another measure of rank correlation whose maximum value 1 is obtained if there is no pair of objects in the same cluster, which is less similar than a pair of objects in different clusters [51]. Like Tau index, the Point Biserial index (**40**) represents the point-biserial correlation coefficient between the pairwise distance matrix and a binary matrix consisting of 0/1 entries that indicate whether or not two objects are in the same cluster [10].

G-plus (**36**) and G-minus (**37**) differ from each other and from the Tau index only in the way ties are treated. The C-Index (**38**) is a normalized sum of the distances between all the pairs of objects that belong to the same cluster; the normalization scheme, which is based on the minimum S_{min} and maximum S_{max} distance sums in the dataset, was proposed in Ref. [49]. The McClain-Rao MCR index (**39**) is the ratio of the average intra-cluster to the average inter-cluster distance [52]. The reciprocal ratio was later (1982) introduced by Good [53] to measure the extent to which clusters are separated and called Index of Separateness.

In 2021, Wiroonsri [25] proposed two indices, here denoted as NC1 (**41**) and NC2 (**42**), with the aim to capture all the potential optimal and sub-optimal partitions for a given dataset to provide the user with more than one solution. Indeed, these cluster validity indices always provide several peaks with different heights, which can be ranked and used to choose the number of clusters that is more appropriate for the specific application. They are based on a correlation measure that is quite similar to the point-biserial correlation (**40**) with the binary entries 0/1 for same/different cluster replaced by the actual distances between the centroids of the clusters where the two objects are located in. More specifically, considering the equation of NC1 (**41**) in Table 3, $\mathbf{d}_X$ is the vector of length n_T collecting the distances between all the pairs of objects in the dataset; $\mathbf{d}_C$ is a vector of the same length with the distances between the corresponding centroids of the clusters the two objects belong to. Then, both vectors used for the calculation of the correlation have a size of $n_T = n{\cdot}(n-1)/2$. We adopted the Pearson correlation coefficient to calculate these indices. NC1 (**41**) and NC2 (**42**) are the proportion and the difference, respectively, of the same two quantities: the first quantity is the normalized correlation increment from K–1 to K clusters while the second quantity is the normalized correlation increment from K to K+1 clusters.

The Dunn's GDI11 index (**43**) is another classical validity index, which dates back to 1973 [54], proposed to identify partitions with compact and well separated clusters. Unlike most of the ratio-type validity indices, Dunn's index has the minimum inter-cluster separation (i.e., the smallest distance between two objects from different clusters or the nearest neighbour distance) in the numerator and the maximum intra-cluster distance, which is sometimes referred to as the cluster diameter and is defined as the largest distance between two objects from

Table 3
Cluster validity indices based on intra- and inter-cluster pairwise distances. The second to last column indicates the applied rule to search for the optimal index value. The quantities N^+, N^-, n_T, n_W, n_B, t, S_W, S_B, S_{min}, S_{max}, $\mathbf{d}_X$, $\mathbf{d}_C$ are explained in the text.

ID	Index (pub. year)	Formula	Rule	Ref.
34	Tau (1974)	$\text{Tau} = \frac{N^+ - N^-}{[(n_T(n_T-1)/2 - t)\cdot(n_T(n_T-1)/2)]^{1/2}} = \frac{N^+ - N^-}{[(n_B \cdot n_W)\cdot(n_T(n_T-1)/2)]^{1/2}}$	first max	[9,50]
35	Gamma (1975)	$\text{GI} \equiv \Gamma = \frac{N^+ - N^-}{N^+ + N^-}$	first max	[9,51]
36	G-plus (1974)	$\text{G+} = \frac{2\cdot N^-}{n_T(n_T-1)}$	first min	[50]
37	G-minus (1974)	$\text{G}- = 1 - \frac{2\cdot N^+}{n_T(n_T-1)}$	first min	[50]
38	C-Index (1976)	$\text{CI} = \frac{S_W - S_{min}}{S_{max} - S_{min}}$	first min	[49]
39	McClain-Rao (1975)	$\text{MCR} = \frac{S_W/n_W}{S_B/n_B}$	max ratio	[52]
40	Point Biserial (1981)	$\text{PB} = \left[\left(\frac{S_B}{n_B} - \frac{S_W}{n_W}\right)\cdot\frac{\sqrt{n_B \cdot n_W}}{n_T}\right] \Big/ s_d$ s_d = standard deviation of all pairwise distances	max	[9]
41	NC1 (2021)	$\text{NC1} = \frac{\text{NC}(K) - \text{NC}(K-1)}{1 - \text{NC}(K-1)} \Big/ \frac{\text{NC}(K+1) - \text{NC}(K)}{1 - \text{NC}(K)}$ $\text{NC}(K) = corr(\mathbf{d}_X, \mathbf{d}_C) \quad K = 2, \ldots, n-1 \quad -1 \le \text{NC}(K) \le +1$ $\text{NC}(1) = 0 \quad \text{NC}(n) = 1$	max	[25]
42	NC2 (2021)	$\text{NC2} = \frac{\text{NC}(K) - \text{NC}(K-1)}{1 - \text{NC}(K-1)} - \frac{\text{NC}(K+1) - \text{NC}(K)}{1 - \text{NC}(K)}$	max	[25]
43–57	Generalized Dunn (1998)	$\text{GDI}pq = \frac{min_{k\neq k'}\delta_{kk'}(p)}{max_k \Delta_k(q)} \quad p = 1,2,\ldots,5 \quad q = 1,2,3$ $\delta_{kk'}(1) = min_{s\in I_k, t\in I_{k'}}\|\mathbf{x}_s - \mathbf{x}_t\|$ $\delta_{kk'}(2) = max_{s\in I_k, t\in I_{k'}}\|\mathbf{x}_s - \mathbf{x}_t\|$ $\delta_{kk'}(3) = \frac{1}{n_k \cdot n_{k'}}\cdot\sum_{s\in I_k}\sum_{t\in I_{k'}}\|\mathbf{x}_s - \mathbf{x}_t\|$ $\delta_{kk'}(4) = \|\mathbf{c}_k - \mathbf{c}_{k'}\|$ $\delta_{kk'}(5) = \frac{1}{n_k + n_{k'}}\cdot\left[\sum_{s\in I_k}\|\mathbf{x}_s - \mathbf{c}_k\| + \sum_{t\in I_{k'}}\|\mathbf{x}_t - \mathbf{c}_{k'}\|\right]$ $\Delta_k(1) = max_{s,t\in I_k}\|\mathbf{x}_s - \mathbf{x}_t\|$ $\Delta_k(2) = \frac{1}{n_k(n_k-1)}\cdot\sum_{s,t\in I_k; s\neq t}\|\mathbf{x}_s - \mathbf{x}_t\|$ $\Delta_k(3) = \frac{2}{n_k}\cdot\sum_{s\in I_k}\|\mathbf{x}_s - \mathbf{c}_k\|$ Note that GDI11 is the original Dunn's index [54]	max	[55]
58	Chou-Su-Lai (2004)	$\text{CSL} = \frac{\sum_{k=1}^{K}\sum_{s\in I_k} max_{t\in I_k}\|\mathbf{x}_s - \mathbf{x}_t\|/n_k}{\sum_{k=1}^{K} min_{k'\neq k}\|\mathbf{c}_k - \mathbf{c}_{k'}\|}$	min	[56]
59	COP (2010)	$\text{COP} = \frac{1}{n}\cdot\sum_{k=1}^{K} n_k \cdot \frac{\sum_{s\in I_k}\|\mathbf{x}_s - \mathbf{c}_k\|/n_k}{min_{t\notin I_k}\, max_{s\in I_k}\|\mathbf{x}_s - \mathbf{x}_t\|}$	min	[57]
60	Silhouette (1987)	$\text{Sil} = \frac{1}{K}\cdot\sum_{k=1}^{K}\frac{1}{n_k}\cdot\sum_{s\in I_k}\frac{b_s - a_s}{max(a_s, b_s)}$ $a_s = \frac{1}{n_k - 1}\cdot\sum_{t\in I_k; t\neq s}\|\mathbf{x}_s - \mathbf{x}_t\| \quad b_s = min_{k'\neq k}\left(\frac{1}{n_{k'}}\cdot\sum_{t\in I_{k'}}\|\mathbf{x}_s - \mathbf{x}_t\|\right)$	max	[42]

(*continued on next page*)

Table 3 (*continued*)

ID	Index (pub. year)	Formula	Rule	Ref.
61	Žalik OS (2011)	$\mathrm{OS} = \dfrac{\sum_{k=1}^{K}\sum_{s\in I_k} o_s}{\sum_{k=1}^{K} min_{k'\neq k}\|\mathbf{c}_{k'} - \mathbf{c}_k\|}$ $\quad o_s = \begin{cases} \dfrac{a_s}{b_s} & if \quad \dfrac{b_s - a_s}{b_s + a_s} < 0.4 \\ 0 & otherwise \end{cases}$ $a_s = \sum_{s,t\in I_k; t\neq s} \|\mathbf{x}_s - \mathbf{x}_t\| \,/\, n_s^{in} \quad \|\mathbf{x}_s - \mathbf{x}_t\| < \|\mathbf{x}_s - \mathbf{c}_k\|$ $b_s = \sum_{s\in I_k; t\notin I_k} \|\mathbf{x}_s - \mathbf{x}_t\| \,/\, n_s^{out} \quad \|\mathbf{x}_s - \mathbf{x}_t\| < \|\mathbf{x}_s - \mathbf{c}_k\|$ n_s^{in} = number of neighbours of the object s in the same cluster n_s^{out} = number of neighbours of the object s in other clusters; if $n_s^{out} = 0$ then $b_s = min_{s\in I_k; t\notin I_k}\|\mathbf{x}_s - \mathbf{x}_t\|$	first min	[41]

the same cluster, in the denominator. Bezdek and Pal [55] later generalized Dunn's index by defining five different measures of distance between clusters and three different measures of cluster diameter. 15 different generalized Dunn indices (**43** to **57**) have been considered in this work, including the original Dunn's index.

Proposed in 1987 by Rousseeuw [42], the silhouette coefficient (**60**) is among the most well-known and widely used validity indices. It is based on the so called silhouette width, which is a measure of the confidence on the membership of each object to its own cluster, obtained from the similarity of the objects in the same cluster compared to other clusters. The silhouette width is the normalized difference between the distance of the object to its 'neighbouring cluster' (i.e., the smallest average distance of the object to the objects belonging to any other cluster) and its average distance to the other objects of the same cluster; it takes values close to 1 when the object lies well within its cluster, values near 0 when it is on the border of two clusters and values close to −1 when it would be more appropriate to be assigned the neighbouring cluster. The silhouette width allows the user to graphically visualize how good a partition is on a point by point and cluster by cluster base.

The Chou-Su_Lai index (**58**) was proposed in 2004 [56] to handle with clusters of different densities and sizes; it is a ratio-type index with the numerator equal to the sum over all the clusters of the average maximum intra-cluster distance (i.e., the measure of cluster cohesion) and the denominator equal to the sum of the minimum inter-cluster separation of each cluster (i.e., the measure of cluster separation). In 2010, based on a similar definition Gurrutxaga et al. introduced the index COP (**59**), where COP stands for Context-independent Optimality and Partiality properties [57]. This index is the weighted mean of ratio-type quantities that characterize each individual cluster, calculated dividing the average distance of the objects in the cluster from the centroid by the distance between the cluster and its nearest object.

Still with the aim to efficiently validate data partitions characterized by the presence of clusters that widely differ in size and density, the index OS (**61**) combines an inter-cluster separation measure with an overlap measure [41]. The separation measure, which is the sum of the

Table 4
Cluster validity indices based on the point-symmetry distance. The second to last column indicates the applied rule to search for the optimal index value.

ID	Index (pub. year)	Formula	Rule	Ref.
62	Sym (2008)	$\mathrm{Sym} = \dfrac{1}{K}\cdot\dfrac{max_{k\neq k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|}{\sum_{k=1}^{K}\sum_{i\in I_k} d_{PS}(\mathbf{x}_i, \mathbf{c}_k)}$	max	[14]
63	Sym-Davies-Bouldin (2009)	$\mathrm{SymDB} = \dfrac{1}{K}\cdot\sum_{k=1}^{K} max_{k'\neq k}\left\{\dfrac{\sum_{s\in k} d_{PS}(\mathbf{x}_s, \mathbf{c}_k)/n_k + \sum_{t\in k'} d_{PS}(\mathbf{x}_t, \mathbf{c}_{k'})/n_{k'}}{\|\mathbf{c}_k - \mathbf{c}_{k'}\|}\right\}$	min	[21]
64	Sym-Dunn (2009)	$\mathrm{SymD} = \dfrac{min_{k\neq k'}\left(min_{s\in I_k, t\in I_{k'}}\|\mathbf{x}_s - \mathbf{x}_t\|\right)}{max_k\left(max_{i\in I_k} d_{PS}(\mathbf{x}_i, \mathbf{c}_k)\right)}$	max	[21]
65	Sym-Generalized Dunn (2009)	$\mathrm{SymGD} = \dfrac{min_{k\neq k'}\left(\dfrac{1}{n_k\cdot n_{k'}}\cdot\sum_{s\in I_k, t\in I_{k'}}\|\mathbf{x}_s - \mathbf{x}_t\|\right)}{max_k\left(\dfrac{2}{n_k}\cdot\sum_{i\in I_k} d_{PS}(\mathbf{x}_i, \mathbf{c}_k)\right)}$	max	[21]
66	Sym-Fukuyama-Sugeno (2009)	$\mathrm{SymFS} = \sum_{k=1}^{K}\sum_{i\in I_k} d_{PS}^2(\mathbf{x}_i, \mathbf{c}_k) - \sum_{k=1}^{K}\|\mathbf{c}_k - \bar{\mathbf{c}}\|^2$	max ratio	[21]
67	Sym-Xie-Beni (2009)	$\mathrm{SymXB} = \dfrac{\sum_{k=1}^{K}\sum_{i\in I_k} d_{PS}^2(\mathbf{x}_i, \mathbf{c}_k)}{n\cdot min_{k\neq k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2}$	min	[21]
68	Sym-Kwon (2009)	$\mathrm{SymKw} = \dfrac{\sum_{k=1}^{K}\sum_{i\in I_k} d_{PS}^2(\mathbf{x}_i, \mathbf{c}_k) + \sum_{k=1}^{K}\|\mathbf{c}_k - \mathbf{b}\|^2/K}{min_{k\neq k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2}$	min	[21]

smallest pairwise distances between cluster centroids, uses all the data objects, while the overlap measure accounts only for the objects that are close to one or more other clusters. The overlap measure is based on the overlap degree of each object, which depends on the average distance of the object to the nearest neighbours in the same cluster and its average distance to the nearest neighbours belonging to other clusters. The overlap degree should account for the cluster shape and can be modulated by the overlap threshold, which was set to 0.4 in this study. Note that, in order to make the calculation of this index feasible for any partition, we set the term b_s equal to the distance of the object s from the first nearest neighbour t of any other cluster in the case the condition $\|\mathbf{x}_s - \mathbf{x}_t\| < \|\mathbf{x}_s - \mathbf{c}_k\|$ was not fulfilled.

2.1.4. Point-symmetry distance indices

The cluster validity indices based on the point-symmetry distance are reported in Table 4. The first one was proposed by Bandyopadhyay and

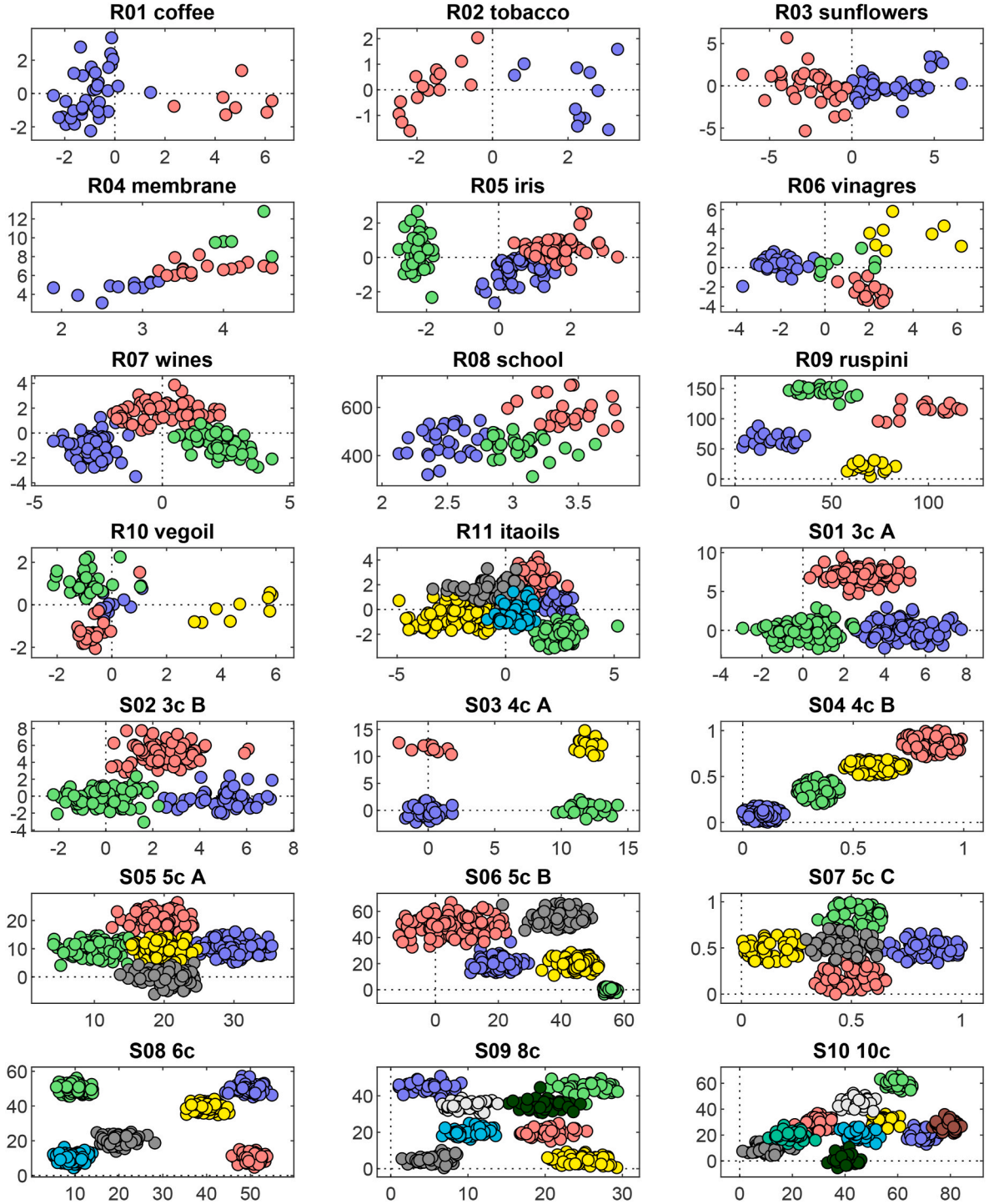


Fig. 1. Two-dimensional scatterplots of the 21 datasets. The data points are coloured according to the groups they belong to in the best partition (i.e., the partition with the maximum adjusted Rand index). For the real datasets, the objects are projected into the space of the first two dimensions of the multidimensional scaling (MDS) analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Saha in 2008 with the name Sym-index [14]. The other indices in this category use the same mathematical definition as some classical CVIs but replace the Euclidean metric with the point-symmetry distance to measure the proximity of the objects to the cluster centroid.

To calculate the point-symmetry distance, let the symmetrical (reflected) point of $\mathbf{x}_i$ with respect to the centroid $\mathbf{c}_k$ of its cluster k be $\mathbf{x}_i^* = 2{\cdot}\mathbf{c}_k - \mathbf{x}_i$. Then, the point-symmetry distance of $\mathbf{x}_i$ from the centroid is:

$$d_{PS}(\mathbf{x}_i, \mathbf{c}_k) = d_{sym}\left(\mathbf{x}_i^*, \mathbf{c}_k\right) \cdot \|\mathbf{x}_i - \mathbf{c}_k\| \tag{10}$$

where the term d_{sym} is the average Euclidean distance of the first two unique nearest neighbours (i.e., $\mathbf{x}_s^{1st}$, $\mathbf{x}_s^{2nd}$) of the symmetrical point of $\mathbf{x}_i$:

$$d_{sym}\left(\mathbf{x}_i^*, \mathbf{c}_k\right) = \frac{\left\|\mathbf{x}_s^{1st} - \mathbf{x}_i^*\right\| + \left\|\mathbf{x}_t^{2nd} - \mathbf{x}_i^*\right\|}{2} \tag{11}$$

Note that the nearest neighbours of $\mathbf{x}_i^*$ are selected only among the objects that are in the same cluster k as the object $\mathbf{x}_i$.

2.2. Datasets

The selection of the datasets is a fundamental step to perform a quantitative comparison of the cluster validity indices. Indeed, for each dataset, it is necessary to have a reliable target value for the expected number of clusters. While for simulated datasets a reliable target is usually handy, except for bizarre cases, for real datasets this is not always the case.

We used 11 real-life datasets that are well-known benchmark datasets for supervised learning applications, for which the object partition into a number of classes is known in advance, and 10 synthetically generated datasets (Fig. 1), which were designed to account for some varying factors, such as the number of clusters (K) from 3 to 10, the dataset size (n) from 100 to 600 objects, the distance between cluster centroids to obtain different cluster overlap, the data distribution within clusters and the presence of noise. The main features of the selected datasets are reported in Table 5. The real-life datasets are available in the UCI repository [58].

Table 5
Description of the datasets. Each dataset is denoted by an alpha-numerical ID label where R is used for the real datasets and S for the simulated datasets. The dataset size is defined by the number of objects (n) and the number of variables (p). The number of classes refers to the "true" partition of data, while the number of expected clusters is derived from the most similar partition to the "true" partition, as measured by the adjusted Rand index.

ID	Dataset name	n	p	Classes	Expected clusters (K^*)	Ref.
R01	Coffee	43	13	2	2	[59]
R02	Tobacco	26	6	2	2	[60]
R03	Sunflowers	70	21	2	2	[61]
R04	Membrane	36	2	3	3	[62]
R05	Iris	150	4	3	3	[63]
R06	Vinagres	66	20	3	4	[64]
R07	Wines	178	13	3	3	[65]
R08	School	45	2	3	3	[66]
R09	Ruspini	75	2	4	4	[67]
R10	Vegoil	83	7	4	4	[68]
R11	Itaoils	572	8	9	6	[69]
S01	3c A	300	2	3	3	This work
S02	3c B	250	2	3	3	This work
S03	4c A	100	2	4	4	This work
S04	4c B	400	2	4	4	This work
S05	5c A	500	2	5	5	This work
S06	5c B	500	2	5	5	This work
S07	5c C	250	2	5	5	This work
S08	6c	600	2	6	6	This work
S09	8c	400	2	8	8	This work
S10	10c	500	2	10	10	This work

2.3. Software

The calculations of all the cluster validity indices and their analysis were performed in MATLAB software, using home-written scripts and appropriate packages. The software Pajek [70] was used to calculate the Minimum Spanning Tree (MST) and the Maximally Regular Graph (MRG).

3. Results

Calculation of cluster validity indices requires that a clustering algorithm be iteratively run over a dataset increasing at each iteration the value of the parameter K, that is, the number of clusters, from a minimum to a maximum value. For each value of K, a different object partition is obtained and a corresponding value of the CVI is computed. Then, the CVI values computed for all the partitions are evaluated to select the optimal CVI value and, accordingly, estimate the number of clusters for the dataset in analysis. This calculated value is finally compared with the expected number of clusters, in order to evaluate the CVI predictive ability.

3.1. Setting a common reference for CVI comparison

The set of partitions used to calculate the validity index has a relevant impact on the CVI values and, hence, on the evaluation of its prediction ability. However, several different partitions of the objects can be obtained for a dataset depending on the specific configuration of the clustering algorithm. Thus, we decided to define a common reference set of partitions for each of the selected datasets to have a fair comparison of the performance of all the indices.

The reference partitions were computed by the k-means clustering algorithm, which is one of the most used clustering methods due to its simple mathematical background and easy implementation. The k-means algorithm determines a partition of the objects into K groups such that the objects within each cluster are more similar to each other than to the objects belonging to the other clusters; the resulting clusters are usually centered in high-density regions of the data space. Similarities among the objects were evaluated by the Euclidean metric after data autoscaling. For each dataset, the set of partitions was determined with the parameter K ranging from 2 to the maximum number of clusters defined as $K^{max} = \sqrt{n}$, where n is the total number of objects in the considered dataset [50].

The major limitation of k-means is the accuracy of the initial location of the random centroids of the clusters, which strongly influences the final partition of a dataset. To avoid this drawback and render the final partition as most reproducible as possible, for each value of the parameter K, the k-means algorithm was repeated 1000 times and the partition with the minimum sum of the within-cluster pairwise distances was retained as the best partition in order to avoid an additional source of data variation in the index comparison.

3.2. Defining the number of expected clusters

Following the comparative methodology proposed by Gurrutxaga et al. [2], to evaluate the performance of CVIs we did not use as the reference the "true" number of clusters (i.e., classes) but the best partition for each dataset, which was defined as the most similar partition to the natural partition of the dataset. This reference partition, which is not always the one with the "true" number of clusters [23], was searched for within the set of all the available partitions of a dataset by using the so-called partition similarity measures.

In particular, to perform this task, the adjusted Rand index r [71] has been calculated to evaluate the congruity of the partition provided by the clustering algorithm with the "true" partition formed by the known classes of the objects. According to the approach of Rand [72] (also known as simple matching or Sokal-Michener index), one counts the number of pairs (a) of objects that are in the same cluster both in the

calculated and in the "true" partition and the number of pairs (d) of objects that are in different clusters in both partitions. The sum of these numbers (i.e., $a + d$) represents the total number of agreements in the comparison and, normalized by the total number of distinct pairs of objects in the dataset, it ranges from around 0 (i.e., dissimilar partitions) to 1 (i.e., identical partitions). This index can be computed on the contingency table that compares the pair assignments made by two partitions. Hubert and Arabie have suggested a modified form (the modified Rand index or adjusted Rand index), which corrects the index for chance as shown below [71].

Let $\mathbf{Q}(R \times C)$ be the contingency table with the number R of rows equal to the number of clusters in the calculated partition and the number C of columns equal to the number of classes of the dataset, 1) q_{ij} denotes the entry of the contingency table, that is, the number of objects that are common to cluster i in the calculated partition and cluster j in the "true" partition; 2) using the standard "dot" notation for row and column sums, $q_{i\cdot}$ and $q_{\cdot j}$ denote the total number of objects in cluster i of the calculated partition and the total number of objects in cluster j of the "true" partition, respectively; 3) n is the total number of objects in the dataset.

Then, the adjusted Rand index r is calculated from the contingency table $\mathbf{Q}$ as [49,73]:

$$r = \frac{(a+d) - Nc}{\frac{n(n-1)}{2} - Nc} \qquad \approx 0 \le r \le 1 \tag{12}$$

where, $(a + d)$ and Nc (i.e., the chance correction term), are defined as:

$$a + d = \frac{n \cdot (n-1)}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} q_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{R} q_{i\cdot}^2 + \sum_{j=1}^{C} q_{\cdot j}^2\right) \tag{13}$$

$$Nc = \frac{n \cdot (n-1)}{2} + \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} q_{i\cdot}^2 \cdot q_{\cdot j}^2}{n \cdot (n-1)} - \frac{1}{2}\left(\sum_{i=1}^{R} q_{i\cdot}^2 + \sum_{j=1}^{C} q_{\cdot j}^2\right) \tag{14}$$

If the relationship between two partitions is comparable to that of partitions picked at random, the adjusted Rand index returns a value close to 0. Small negative values can be obtained for cases where the partition agreement is less than expected by chance. In all the datasets in analysis and for all the calculated partitions, the adjusted Rand index was greater than 0.2; this value was proposed as threshold for random partitions. The partition with the maximum value of this index was selected as the best partition for the dataset and the number of clusters in that partition was taken as the target for the CVI quality evaluation. Table 5 collects the expected number of clusters for all the datasets, which was defined according to this approach. The target value for the parameter K coincides with the number of known classes in the "true" partition for all the datasets with only two exceptions: Vinagres and Itaoils. For the dataset Vinagres (R06), which has 3 classes of objects, the expected number of cluster is 4; this is a quite reasonable result considering that this dataset shows four main high-density regions in the data space (Fig. 1). The same consideration holds for the dataset Itaoils (R11), for which the data structure is complex with a large overlap between classes.

3.3. Analysing the performance of cluster validity indices

Although the partition sets were specifically determined by the k-means algorithm, the validity indices surveyed in the present study are quite general and can be adopted to estimate the number of clusters for any clustering method. Indeed, we preferred to examine only those indices that were method independent. Moreover, the indices requiring external information or tuning parameters were not considered. Finally, we selected those indices that are formulated in such a way an automatic decision rule can be used to provide an objective prediction and, hence, avoid the problem of human subjectivity. The adopted rules to search for the optimal index value and, accordingly, the optimal number of clusters for each dataset are reported in the same tables as the mathematical definitions of the indices (Tables 1–4). Fig. 2 shows the characteristic behaviours of some indices for each type of adopted decision rule.

Before index calculation, data were autoscaled to avoid the influence of the different measurement scales of the variables on the similarity metrics, as it often happens in real datasets. For the majority of the CVIs, the similarity between objects and cluster centroids was evaluated by the Euclidean metric.

Each index was allowed to adopt the most favourable conditions to optimize its performance. In addition, the index behaviour was analysed by increasing the number of clusters up to the maximum value equal to the square root of the total number of objects in the dataset. This choice stems from the consideration that some indices exhibited degenerative behaviour when the number of clusters approaches the number of objects in the dataset and such behaviour may lead to misleading predictions. For the same reason, in some cases, the decision rules 'the first minimum' and 'the first maximum' were preferred and adopted instead of the absolute minimum and maximum of the index. The calculated

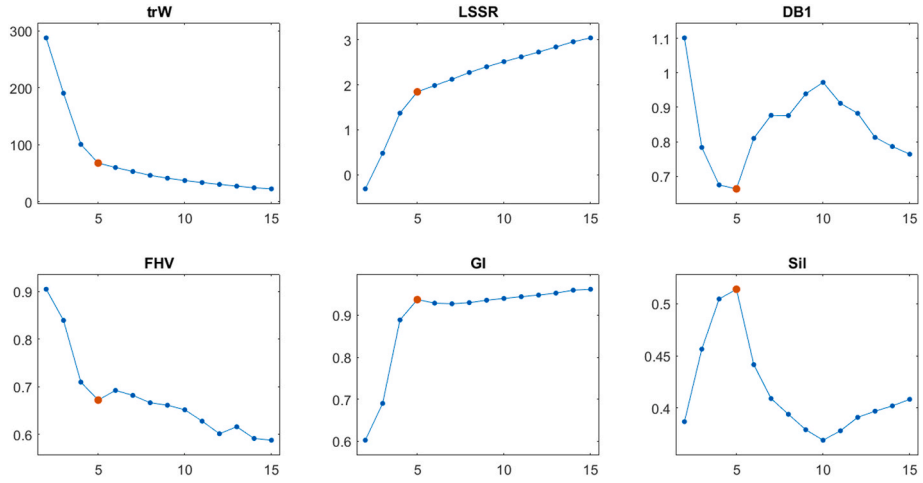


Fig. 2. Examples of six different behaviours of CVIs for the simulated dataset S07 with five expected clusters: CVI value (vertical axis) *vs* number K of clusters (horizontal axis). The red dot indicates the estimated number of clusters. The corresponding rules to search for the optimal K parameter are: 'max ratio' for the indices trace_W (trW) and Hartigan (LSSR); absolute 'min' for Davies-Bouldin index (DB1); 'first min' for FuzzyHyperVolume index (FHV); 'first max' for the Gamma index (GI); absolute 'max' for Silhoutte index (Sil). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

values of all the CVIs for each dataset are provided as supplementary material both in the tabular (Table S1) and graphical format (Fig. S1).

3.4. Correlations between cluster validity indices

In order to disclose the main linear relationships among the CVIs, we calculated the Pearson correlation coefficient between all the pairs of indices for each dataset to evaluate which indices have similar behaviour and, hence, may predict the same number of clusters. The complete correlation matrix, which was obtained by averaging the correlation coefficients over all the datasets, is provided in the supplementary material (Table S2), while Table 6 highlights the relevant average correlations between the pairs of highly correlated indices.

The RS index (**8**) is correlated −1 with trW (**1**) due to the constraint $TSS = WSS + BSS$, that is, the total sum of squares, which is constant for a given dataset, depends both on the within-group sum of squares and the between-group sum of squares. Then, the index RS can be defined as function of trW as the following:

$$\mathrm{RS} = \frac{BSS}{TSS} = \frac{TSS - WSS}{TSS} = 1 - \frac{WSS}{TSS} = 1 - \frac{\mathrm{trW}}{TSS} \tag{15}$$

Moreover, Fukuyama-Sugeno FS index (**16**), Ball-Hall BH (**2**), Banfield-Raftery BR (**3**), and Hartigan LSSR (**6**) are strongly correlated each other and with trW (**1**) since all of them are different functions of the within-group sum of squares. The large average correlation (0.975) observed between the index WCH (**5**) and Calinski-Harabasz CH (**4**) is due to the lack of relevant overlapping between clusters in several datasets.

As expected, the modified Wemmert-Gancarski WG2 index (**14**) has high average correlation (0.958) with the Silhouette Sil index (**60**), WG2 being conceptually conceived with the same underlying idea as Silhouette.

The correlation between Xie-Beni XB1 index (**17**) and Kwon Kw index (**19**) is always greater than 0.960. Indeed, the additional quantity present in the Kwon index has the following limit:

$$lim_{K\to n} \frac{1}{K} \cdot \sum_{k=1}^{K} \|\mathbf{c}_k - \mathbf{b}\|^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{b}\|^2 = \frac{TSS}{n} \tag{16}$$

which is a constant. The Kwon index also has a large average correlation of 0.953 with Tang Tn index (**20**).

Correlations greater than 0.950 are also present in a subset of generalized Dunn's indices, in particular between the GDIp2 and GDIp3 (p = 1, 2, 3) subsets. Other relevant inverse correlations (<−0.950) have been observed between the pair of CSL (**58**) and SV (**12**), and the Gamma index GI (**35**) with G+ (**36**) and CI (**38**).

3.5. Values of CVIs and prediction of the number of clusters

Through the analysis of the trend of the validity index and applying the selected decision rule, we obtained the optimal value of the index and the corresponding best partition for each dataset, that is, the optimal number of clusters in the dataset. This estimate for the parameter K is reported for all the surveyed CVIs and the 21 datasets in the supporting material (Table S1), along with the calculated values of each validity index for each dataset (Table S2 and Fig. S1).

Some synthetic indices were calculated to measure the overall quality of the CVIs. They are reported in Table 7. The first quality index is the overall score, denoted as N(=), that is, the number of datasets for which the index prediction matches the expected number of clusters or, in other words, the number of correct estimates. It is interesting to note the failures of some indices to calculate the correct number of clusters in those datasets that have well-distinct clusters and error-free cluster structure, such as Ruspini (R07). If an index fails in this straightforward case, then it is unlike that it would provide reliable predictions in more complex real clustering applications. These indices are WCH (**5**), RL (**7**), Vsv2 (**26**), KDW (**31**), FHV (**32**), NI (**33**), G– (**37**), GDI51 (**55**), GDI52 (**56**), GDI53 (**57**), Sym (**62**) and SymD (**64**).

It can also be noted that for all the simulated datasets, the correct number of clusters has been always obtained by the indices trW (**1**), BR (**3**), LSSR (**6**), RS (**8**), DB2 (**10**), PBM (**11**), and WG1 (**13**). Moreover, for the same datasets, the indices CH (**4**), DB1 (**9**), WG2 (**14**), XB2 (**18**), GDI33 (**51**), GDI43 (**54**), COP (**59**), MCR (**39**), Sym (**62**) behave quite satisfactory, providing the expected number of clusters for almost all the simulated datsets with only a difference of one cluster in no more than one dataset. Among all these indices, only BR (**3**) and MCR (**39**) have a generally low performance, due to their problematic behaviour on the real datasets.

Along with the number of successes, we also calculated the number of datasets (denoted by N<) for which the index provides a value of K smaller than the expected value and, on the opposite, the number of datasets (denoted by N>) for which the estimate exceeds the correct number of clusters. Some indices resulted to be particularly sensitive to the presence of noise in the dataset overestimating the correct number of clusters; this is the case of the indices BH (**2**) and Sym-Dunn (**6**), which, for instance, calculated 14 and 13 clusters, respectively, instead of three clusters for the dataset S02.

The quality of the CVIs has also been evaluated considering how big the prediction error is, which was quantified in terms of mean absolute deviation as:

$$DS_i = \frac{\sum_{m=1}^{M} \Delta_{im}}{M} \qquad m = 1, M \tag{17}$$

where the summation runs over all the datasets (i.e., $M = 21$ in this study) and accounts for the absolute difference between the number of clusters $\widehat{K}_{im}$ predicted by the ith index for the mth dataset and the expected number K_m^*:

$$\Delta_{im} = \left| \widehat{K}_{im} - K_m^* \right| \tag{18}$$

By definition, low values of this score are associated to CVIs with the best overall performance. In Table 7, the CVIs are ranked according to this score. Then, it can be easily noted that 11 indices are located in the top of the list with a *DS* score lower than 0.6. Among these, there are some of the traditional CVIs such as trace_W (**1**), Hartigan LSSR (**6**), Davies-Bouldin (**9**), the generalized Dunn's index GDI33 (**51**) and Fukuyama-Sugeno

Table 6
Relevant correlations between pairs of cluster validity indices. Correlation coefficients larger than 0.9 and smaller than −0.9 are highlighted in italics.

Index	BH (2)	BR (3)	LSSR (6)	RL (7)	FS (16)	trWB (27)	SS (29)	LDR (30)	SymFS (66)
trW (1)	*0.994*	*0.951*	*−0.966*	0.837	*0.995*	−0.799	0.889	*−0.931*	*0.920*
BH (2)		*0.947*	*−0.959*	0.839	*0.992*	−0.798	0.889	*−0.927*	*0.915*
BR (3)			*−0.994*	*0.949*	*0.970*	*−0.929*	*0.962*	*−0.992*	*0.965*
LSSR (6)				*−0.931*	*−0.980*	*0.912*	*−0.951*	*0.988*	*−0.963*
RL (7)					0.872	*−0.957*	*0.936*	*−0.961*	*0.916*
FS (16)						−0.841	*0.924*	*−0.955*	*0.942*
trWB (27)							*−0.929*	*0.950*	*−0.915*
SS (29)								*−0.964*	*0.934*
LDR (30)									*−0.961*

Table 7
Quality scores used for the index comparison. The CVIs are ranked according to their overall performance as quantified by the mean absolute deviation score *DS*. N(=): number of datasets for which the calculated number of clusters equals the expected value; N(<): number of underestimates; N(>): number of overestimates; μ: sensitivity index (μ = 0 indicates invariance to the total number of allowed partitions).

Index ID	Name	Symbol	*DS*	N(=)	N(<)	N(>)	μ	Rule
6	Hartigan	LSSR	0.381	14	4	3	0.13	max ratio
1	Trace_W	trW	0.429	15	2	4	0.13	max ratio
8	RS	RS	0.429	15	2	4	0.13	max ratio
59	COP	COP	0.429	14	6	1	0	min
10	Davies-Bouldin*	DB2	0.476	15	3	3	0.10	min
13	Wemmert-Gancarski	WG1	0.476	15	4	2	0.06	max
9	Davies-Bouldin	DB1	0.524	14	4	3	0.25	min
16	Fukuyama-Sugeno	FS	0.524	15	6	0	0.41	max ratio
51	Generalized Dunn 33	GDI33	0.524	14	6	1	0.13	max
54	Generalized Dunn 43	GDI43	0.524	14	5	2	0.10	max
18	Xie-Beni*	XB2	0.571	13	6	2	0.13	min
4	Calinski-Harabasz	CH	0.619	14	5	2	0.06	first max
11	Pakhira-Bandyopadhyay-Maulik	PBM	0.619	14	3	4	0.03	max
20	Tang	Tn	0.619	13	7	1	0	min
19	Kwon	Kw	0.667	12	8	1	0	min
21	Partition Separation	PS	0.714	14	5	2	0.06	max
25	Kim-Park v_{sv}	Vsv1	0.714	12	2	7	0.19	first min
17	Xie-Beni	XB1	0.762	12	7	2	0.03	min
14	Wemmert-Gancarski*	WG2	0.810	13	4	4	0.25	max
23	Kim-Ramakrishna SD*	SD2	0.857	11	8	2	0.32	min
30	Scott-Symons 2	LDR	0.857	10	11	0	0	max ratio
50	Generalized Dunn 32	GDI32	0.857	10	10	1	0	max
53	Generalized Dunn 42	GDI42	0.857	12	5	4	0.25	max
60	Silhouette	Sil	0.857	11	10	0	0	max
47	Generalized Dunn 22	GDI22	0.905	11	10	0	0	max
48	Generalized Dunn 23	GDI23	0.905	14	2	5	0.35	max
65	Sym-Generalized-Dunn	SymGD	1.000	10	8	3	0.22	max
22	Rezaee-Lelieveldt-Reiber	SD1	1.048	10	8	3	0.06	min
26	Kim-Ramakrishna v_{sv}*	Vsv2	1.048	11	7	3	0.29	first min
39	McClain-Rao	MCR	1.095	9	10	2	0	max ratio
68	Sym-Kwon	SymKw	1.095	12	1	8	0.63	min
5	WCH	WCH	1.190	11	5	5	0.13	max
52	Generalized Dunn 41	GDI41	1.190	10	8	3	0.10	max
62	Sym	Sym	1.190	12	4	5	0.38	max
38	C-Index	CI	1.238	12	3	6	0.13	first min
67	Sym-Xie-Beni	SymXB	1.238	10	7	4	0.10	min
41	NC1	NC1	1.286	10	6	5	0.60	max
12	Zalik SV	SV	1.333	11	8	2	0.06	first max
32	Fuzzy HyperVolume	FHV	1.333	11	2	8	0.13	first min
40	Point Biserial	PB	1.333	7	11	3	0	max
44	Generalized Dunn 12	GDI12	1.381	11	7	3	0.29	max
34	Tau	Tau	1.429	7	11	3	0.03	first max
35	Gamma	GI	1.429	11	3	7	0.13	first max
46	Generalized Dunn 21	GDI21	1.429	7	10	4	0	max
24	Halkidi	SDbw	1.524	12	3	6	0.32	first min
45	Generalized Dunn 13	GDI13	1.524	10	5	6	0.32	max
29	Scott-Symons 1	SS	1.571	13	0	8	0.06	max ratio
42	NC2	NC2	1.571	10	11	0	0.10	max
49	Generalized Dunn 31	GDI31	1.571	9	8	4	0.22	max
3	Banfield-Raftery	BR	1.619	9	12	0	0	max ratio
15	Score Function	SF	1.619	12	1	8	0.57	max
61	Zalik OS	OS	1.762	7	11	3	0.06	first min
37	G-minus	G–	1.810	5	15	1	0	first min
43	Generalized Dunn 11	GDI11	1.810	9	7	5	0.19	max
58	Chou-Su-Lai	CSL	1.952	10	5	6	0.22	min
56	Generalized Dunn 52	GDI52	2.000	11	3	7	0.48	max
57	Generalized Dunn 53	GDI53	2.000	6	15	0	0	max
7	Ratkowsky-Lance	RL	2.048	7	14	0	0	max ratio
63	Sym-Davies-Bouldin	SymDB	2.095	1	11	9	0.03	min
28	Friedman-Rubin 2	DR	2.143	11	3	7	0.60	max ratio
36	G-plus	G+	2.333	9	5	7	0.10	first min
66	Sym-Fukuyama-Sugeno	SymFS	2.571	10	3	8	0.32	max ratio
33	Negentropy Increment	NI	2.619	8	5	8	1.02	first min
27	Friedman-Rubin 1	trWB	2.667	8	2	11	0.41	max ratio
64	Sym-Dunn	SymD	2.857	7	4	10	0.54	max
2	Ball-Hall	BH	3.000	9	1	11	1.43	max ratio
55	Generalized Dunn 51	GDI51	3.143	1	14	6	0.03	max
31	Marriot	KDW	3.476	6	6	9	0.51	first min

(**16**). More recent indices in the top list are those proposed by Kim et al. [16] as variants of Davies-Bouldin index, that is, DB2 (**10**), and Xie-Beni index, that is, XB2 (**18**), along with the index COP (**59**) proposed by Gurrutxaga et al. [57] and Wemmert-Gancarski WG1 index (**13**), which has been described by Desgraupes [4]. On the opposite side, a large block of indices is located at the bottom of the list due to an overall unsatisfactory performance. Fig. 3 shows, for the real and simulated datasets, the distribution of the differences between the *i*th calculated number of clusters $\widehat{K}_{im}$ and the expected number K_m^* for each dataset. The median of the estimation errors is generally zero indicating that most of the CVIs correctly provide the expected number of clusters in the dataset. Among the simulated datasets, an exception is the dataset S08 for which most of the indices predict 4 instead of 6 clusters, thus providing a prediction error equal to −2. Among the real datasets, positive medians are observed for the datasets Sunflowers R03, Membrane R04, and Vinagres R06, for which on average more clusters than expected are predicted, while negative medians are observed for the datasets Iris R05, School R08 and Itaoils R11, for which on average less clusters are estimated. The datasets with more relevant prediction errors are those with a more complex data structure. For instance, most of the CVIs fail to estimate the correct number of clusters for the dataset Itaoils R11, giving on average an underestimate of the number of clusters. There are a few indices that provide a very large number of clusters for this dataset likely due to their degenerative behaviour as the parameter *K* increases, e.g., KDW (**31**), NI (**33**), BH (**2**), SymDB (**63**), BR (**3**), trWB (**27**), DR (**28**). Some of these

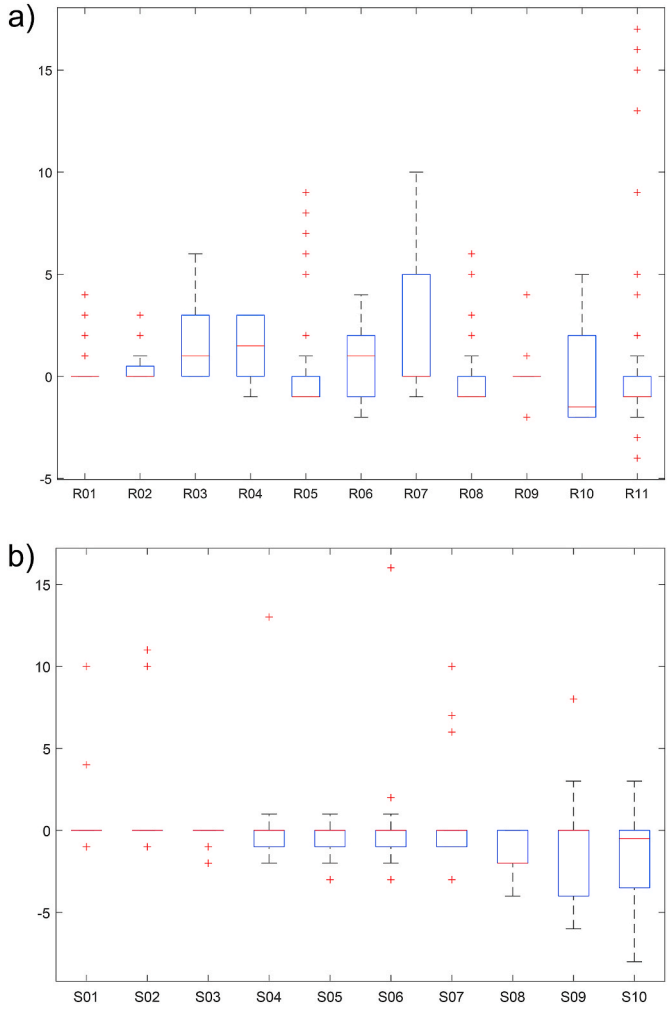


Fig. 3. Box-plots of the CVI estimation errors (i.e., differences between the calculated $\widehat{K}$ and the expected number K^* of clusters) for the a) real datasets and b) simulated datasets.

indices are based on the within-group scatter matrix, which cannot be properly calculated in the case of clusters with a small number of objects. It is noteworthy that among the few indices able to predict the expected number of clusters for the dataset Itaoils R11 there are the FuzzyHyperVolume index (**32**), the C-Index (**38**), the original Dunn's index GDI11 (**43**) and all the generalized Dunn indices with the first and third cluster separation measure, along with their two variants based on the point-symmetry distance, that is, SymD (**64**) and SymGD (**65**).

3.6. Sensitivity analysis of the cluster validity indices

The validity indices can be calculated for each value of the parameter *K* up to its maximum value that naturally coincides with the number *n* of objects in the dataset. However, the index calculation is usually limited to a maximum value K^{max}, which is smaller than *n* and is *a-priori* decided by the user. A simple rule of thumb has been proposed as $K^{max} = \sqrt{n}$, where *n* is the total number of objects [20,23,74]. In some cases, it has been suggested to select a general low value (e.g. 10) for the maximum number of clusters, regardless of the total number of objects in the dataset [1,19]. Since the calculation of the correct number of clusters depends on the index behaviour while increasing the parameter *K*, the choice of the maximum value of *K* may influence the index estimation. This is especially true for those indices that have more than one minimum or maximum and for the indices with degenerative trend after a certain value of *K*. Therefore, to test the robustness of the CVI prediction, we performed a sensitivity analysis by varying K^{max} around the default value (i.e., the square root of the number of objects), that is, $K^{max} = \sqrt{n} - 1$ and $K^{max} = \sqrt{n} + 1$.

The comparison of the results provided by a validity index for the three different values of K^{max} has been performed by calculating the mutual variability μ as:

$$\mu = \frac{|\widehat{K}_{-1} - \widehat{K}| + |\widehat{K}_{-1} - \widehat{K}_{+1}| + |\widehat{K} - \widehat{K}_{+1}|}{3} \tag{19}$$

where $\widehat{K}$, $\widehat{K}_{-1}$ and $\widehat{K}_{+1}$ indicate the predicted number of clusters in the case of $K^{max} = \sqrt{n}$, $K^{max} = \sqrt{n} - 1$ and $K^{max} = \sqrt{n} + 1$, respectively. The mean mutual variability on the 21 datasets is reported for each CVI in Table 7. Only 19% of the validity indices resulted invariant ($\mu = 0$) to the variation of the maximum value of the parameter *K* even if this conclusion should be taken with caution since only a small variation around the default value has been considered.

Moreover, for the indices whose performance is more severely influenced by the maximum possible number K^{max} of clusters, a deeper analysis has been performed by comparing the *DS* scores obtained for the different levels of K^{max}. In all the cases, the best predictions have been obtained setting the maximum value of the parameter *K* equal to $\sqrt{n} - 1$. This rule seems to be more appropriate for all the datasets with a relatively small number of objects. Moreover, a reduced interval of the *K* values limits unreliable predictions due to possible degenerative behaviours of some indices when the value of *K* approaches the number of dataset objects.

3.7. Multivariate comparison of CVIs

Principal Component Analysis (PCA) and the Minimum Spanning Tree (MST) approach were applied to the results of the 68 validity indices for all the 21 considered datasets to allow an easier overall comparison of CVIs and for a deep comprehension of their mutual relationships. This multivariate analysis has been carried out on the data matrix (68 × 21) collecting the performance of each CVI for each dataset as measured by the absolute prediction error, that is, the absolute difference between the optimal number of clusters provided by the CVI and the expected number for each dataset. To run PCA, no scaling was applied to the data matrix. For MST, the Manhattan metric was selected

to calculate the similarities between all the pairs of CVIs. Moreover, a theoretical validity index, called *Best* and denoted by the symbol B, has been added to the set of CVIs. This optimal reference corresponds to a theoretical index which is able to predict the correct number of clusters for all the datasets and therefore, its absolute prediction error is always equal to zero.

The results of PCA are shown in Fig. 4. In the score plot, the validity indices are coloured according to their overall performance as quantified by the score *DS* (i.e., the mean absolute prediction error reported in Table 7): the indices with the best overall performance are in green, the worst indices are in red and the indices with average overall performance are in grey. The first principal component (PC1), which explains around 50% of the total variance, ranks the CVIs from the best (on the right) to the worst ones (on the left). The second principal component PC2, with around 13% of explained variance, highlights the differences among the less performing indices and, in particular, distinguishes between the two different types of decision error that can occur. The indices with large value of PC2 (e.g., SymDB, DR, trWB, BR, BH, NI and KDW) generally tend to overestimate the parameter *K* and, hence, to predict a data partition with too many clusters. In particular, these indices fail to predict the correct number of clusters for the most complex datasets with relevant overlap among the clusters (e.g. Wines R07 and Itaoils R11). On the contrary, the indices with lower PC2 score tend to indicate fewer clusters than the clusters actually present in the natural data partition. This second type of error might be considered more serious in most practical applications of cluster analysis due to the information lost when merging distinct clusters [10]. It is noteworthy that among these indices with less satisfactory results there are the three generalized Dunn's indices defined with the fifth type of cluster separation measure (i.e., GDI51, GDI52, GDI53), which takes into account the minimum average joint distance of

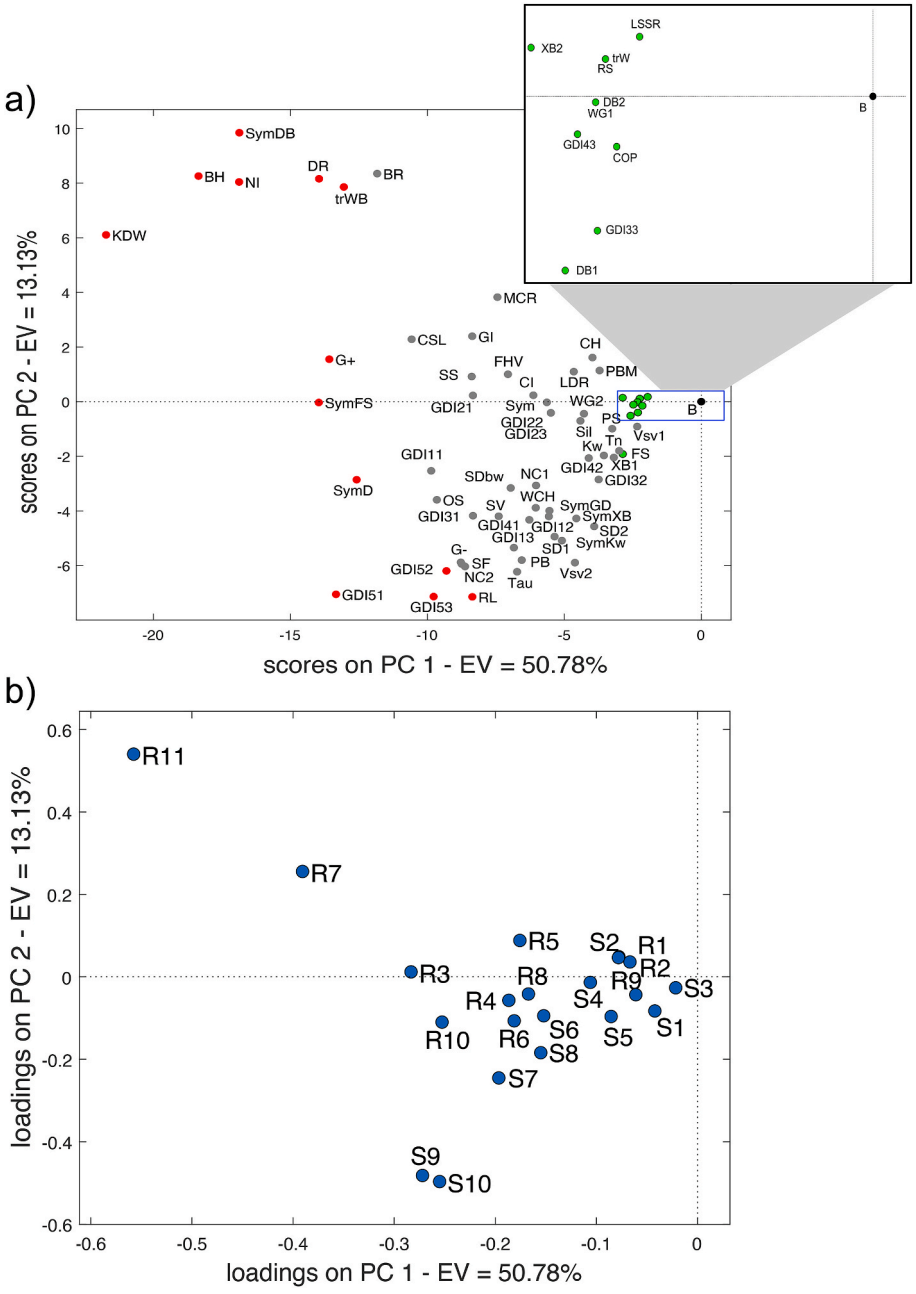


Fig. 4. PCA a) score and b) loading plot of the first two principal components. CVIs are coloured according to the *DS* quality score: the top-ranked indices are in green; the worst indices are in red; the indices with average overall performance are in grey. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

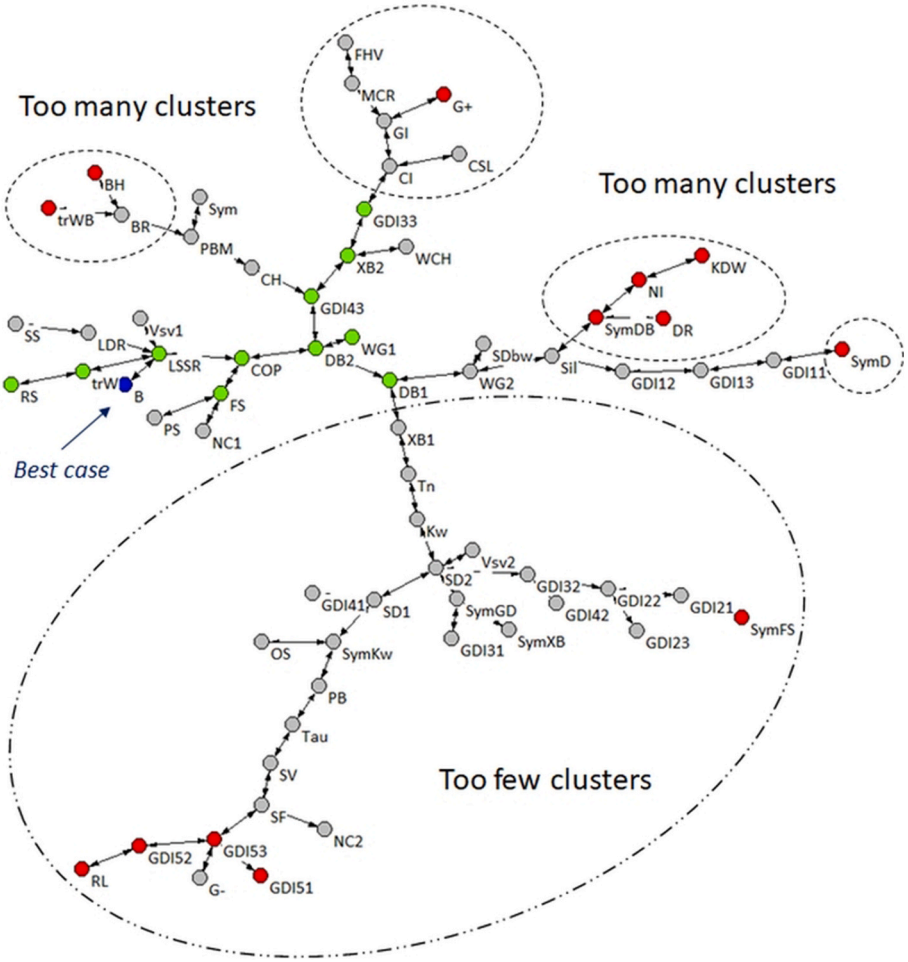


Fig. 5. Minimum Spanning Tree (MST) by Manhattan metric on the differences between calculated and expected number of clusters for the 21 datasets. CVIs are coloured according to the *DS* quality score: the top-ranked indices are in green; the worst indices are in red; the indices with average overall performance are in grey. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the objects of two different clusters from their centroids. Their prediction errors are more relevant for the simulated datasets S09 and S10 with eight and ten expected clusters, respectively.

The relationships among the studied cluster validity indices were further investigated by the MST approach. MST has been performed on the same data matrix as PCA using the Manhattan metric to calculate the similarities between all the pairs of indices (Fig. 5). The advantage of MST is the data visualization in the form of a tree graph, where similar indices are located in the same tree branch. The most dissimilar indices are the terminal nodes of the tree branches, while the indices with more mutual relationships are represented by the most branched nodes. Although in a different way, the tree structure of MST analysis is in accordance with the PCA score plot. However, the CVIs that generally overestimate the number of clusters are partitioned into three different subgroups, according to their different analogies among the prediction errors. The optimality region gathers the best indices already highlighted in the PCA score plot, but also taking into account their reciprocal similarities.

3.8. Application of CVIs to complex data

As the final evaluation step of the CVIs, we tested their performance on data with complex structure. To this aim, we selected a joint metabolomics dataset being comprised of fluorescence spectroscopy, ^{1}H NMR spectroscopy (CPMG and NOESY-Presat) and biomarker measurements (TIMP-1 and CEA) on human plasma from cancer and control samples [75]. The dataset includes 94 samples described by 476 variables, which were properly pre-processed in the original study in order to avoid the common scaling problem in multiblock modelling. The first two variables are the biomarkers, the next 19 are the fluorescence data as PARAFAC scores and the last 455 are the NMR peaks.

Unlike the benchmark datasets used in the previous comparative study, which were generally characterized by a small number of

Table 8
Estimated number of clusters for the metabolomics dataset according to the different CVIs. Indices with rank up to 20 on *DS* score, excluding those with the max ratio rule, are highlighted in boldface.

Estimated $\widehat{K}$	Cluster Validity Indices
2	**CH** (4), WCH (5), SF (15), **Vsv1** (25), Vsv2 (26), DR (28), Tau (34), GI (35), G+ (36), G– (37), CI (38), PB (40), NC2 (42), GDI11 (43), GDI12 (44), GDI13 (45), GDI21 (46), GDI22 (47), GDI23 (48), GDI31 (49), GDI32 (50), **GDI33** (51), GDI51 (55), GDI52 (56), GDI53 (57), **COP** (59), SymD (64), SymGD (65)
3	**PS** (21)
4	LSSR (6), NC1 (41)
5	SD1 (22), **SD2** (23), SS (29)
6	BH (2), BR (3), RL (7), **PBM** (11), FHV (32), NI (33), Sil (60)
7	trW (1), RS (8), FS (16), trWB (27), LDR (30), Sym (62), SymFS (66)
8	**DB1**(9), **DB2** (10), SV (12), **WG2** (14), SDbw (24), MCR (39), CSL (58), SymDB (63)
9	**XB1** (17), **XB2** (18), **Kw** (19), **Tn** (20), GDI42 (53), **GDI43** (54), SymXB (67), SymKw (68)
10	**WG1** (13), KDW (31), GDI41 (52), OS (61)

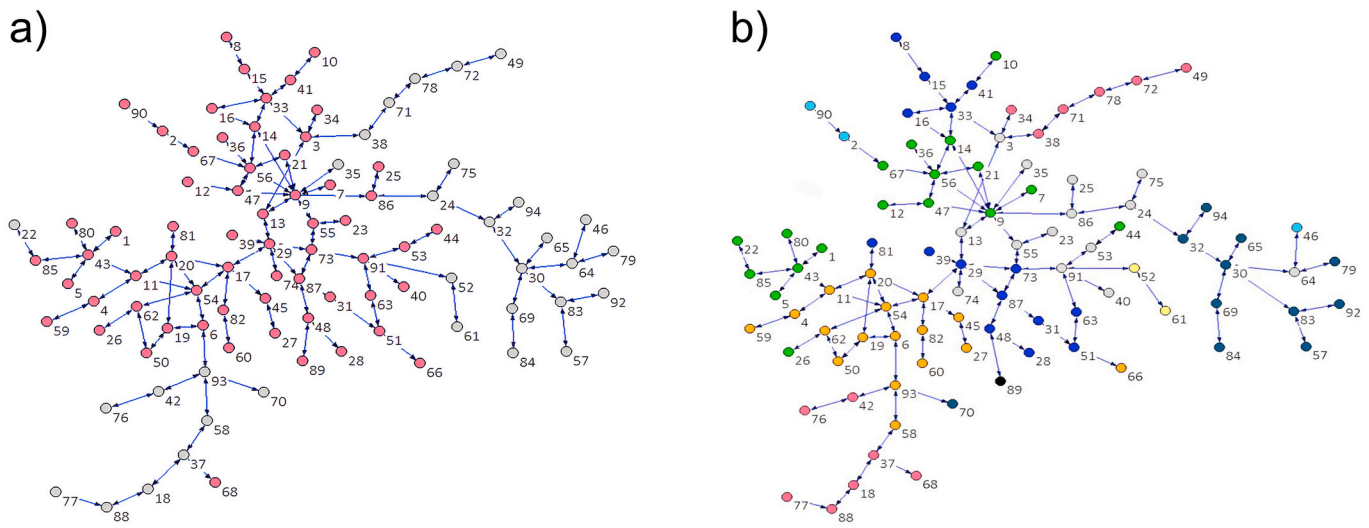


Fig. 6. Graphical representation of the metabolomics dataset by Maximally Regular Graph (MRG) on Euclidean pairwise distances between samples. The graph vertices correspond to the human blood samples, which are coloured according to the cluster they belong to; the graph edges represent their similarity relationships: a) sample partition in 2 clusters; b) sample partition in 9 clusters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

variables, this dataset better represents the most common real-life scenarios of complex data that require clustering algorithms to elucidate the unknown data structure in presence of high number of variables.

Since the number of variables is higher than the number of samples, a variable reduction has been necessary to allow the calculation of all the indices; in effect, the CVIs based on the calculation of the scatter matrix determinant cannot be computed when there is an excess of variables, as for the indices listed in Table 2. Moreover, some variables may be redundant and/or noisy leading to undesired effects on the measures of the similarity relationships between samples. Therefore, we performed a Principal Component Analysis (PCA) on the autoscaled variables and finally retained the first 34 PCs with eigenvalue larger than 1. The reference partitions were therefore computed on the final data matrix of dimension 94×34 by the *k*-means clustering algorithm; similarities between objects were evaluated by the Euclidean metric with no further scaling of the variables to preserve the variance explained by each PC and the parameter K was tuned from 2 to the maximum number of clusters chosen equal to $K^{max} = int(\sqrt{94}) = 10$. The adjusted Rand index was very low (around 10^{-2}) for all the calculated partitions, meaning that there was no congruity between the calculated partitions by *k*-means algorithm and the known partition of cancer and control samples. For this reason, the metabolomics dataset was not used in the general comparative study of CVIs since it lacks a reliable target for the expected number of clusters, which is required to calculate the quality scores used to compare the validity indices. Nonetheless, the results on this dataset provided further insights into the functioning and performance evaluation of CVIs when applied to data with complex structure. Table 8 shows the estimated number of clusters by each validity index according to the automated selection rule as defined in Table 7. The graphical visualization of the most frequent partitions, that is, the partitions obtained by several CVIs, are provided in Fig. 6 by the Maximally Regular Graph (MRG), which is a graph representation of the similarity relationships among the samples; MRG is a variant of the Minimum Spanning Tree, which is generated optimizing the graph complexity by adding back to the original MST, one by one, the missing connections previously skipped during the computation of the MST itself [76].

Most of the indices provided the 2-cluster partition as the best one (Fig. 6a); however, also the partition with 9 (Fig. 6b) clusters is noteworthy since it was selected by several indices that are among the best performing indices in the comparative study on the benchmark datasets (these indices are highlighted in boldface in Table 8). From a deeper exploration of these partitions, it was concluded that while the 2-cluster partition roughly divide the samples in two heterogeneous groups, which do not correspond to the cancer and control sample groups, the many-cluster partitions seem to be more interesting since they provide subgroups of more homogenous samples and some apparent outliers. In this case, there is no optimal partition but a number of possible partitions that need to be evaluated on the basis of additional expert knowledge.

4. Conclusions

In this paper, 68 cluster validity indices (CVIs) have been surveyed and evaluated by comparison on 21 benchmark datasets for which the "true" number of clusters was previously known. Some indices showed an overall good performance providing the expected number of clusters for almost all the datasets in analysis. This group of best CVIs includes Harting index LSSR (**6**), Trace_W index trW (**1**), R-Squared index RS (**8**), COP (**59**), the two Davies-Bouldin indices DB1 (**9**) and DB2 (**10**), Wemmert-Gancarski index WG1 (**13**), Fukuyama-Sugeno index FS (**16**) and the two Generalized Dunn indices GDI33 (**51**) and GDI43 (**54**).

Other indices, such as the modified Xie-Beni index XB2 (**18**), Calinski-Harabasz index CH (**4**), Pakhira-Bandyopadhyay-Maulik index PBM (**11**), Kwon index Kw (**19**), Tang index Tn (**20**), Partition Separation PS (**21**), Kim-Park index Vsv1 (**25**) and the here proposed modified Wemmert-Gancarski index WG2 (**14**) showed a lower but still acceptable overall performance. Also Xie-Beni index XB1 (**17**) and Silhouette (**60**), which are among the most traditional cluster validity indices, behave quite well for all the datasets. In particular, it is noteworthy that Silhouette has a very clear trend for all the datasets without fluctuations, thus providing an unambiguous optimal point determination (Fig. S1).

Despite the ranking based on the overall quality index *DS*, it should be considered that the cluster validity indices that use as the decision rule the absolute or first min/max of the index, such as COP (**59**), DB1 (**9**), DB2 (**10**) and WG1 (**13**), are in principle more reliable in detecting the "natural" number of clusters than those indices based on the max ratio rule, such as LSSR (**6**), trW (**1**) and FS (**16**), since the detection of the most relevant change in monotonic distributions is not always an easy task and involves a certain amount of subjectivity. Indeed, the estimated number of clusters for the indices based on monotonic functions strongly depends

also on the extrapolated values at $K = 1$ and $K = K^{max} + 1$, and is thus susceptible to significant variation for small variation of the considered partition. Moreover, especially for small datasets, their behaviour tends to be quite similar to a straight line; in this case, the detection of the optimal number of clusters is theoretically unfeasible.

With respect to the invariance to changes of the maximum allowed number of clusters, 14 out of 68 indices (21%) are fully invariant ($\mu = 0$); 9 of them are based on the max rule, 4 on the min rule and 1 on the first min rule. Other 12 indices show a sensitivity less than 0.1 ($0 < \mu < 0.1$), and, among these, only two indices are based on the max ratio rule, namely BR (**3**) and RL (**7**). In quantitative terms, only 15% of the indices based on the max ratio rule is almost invariant to the change of the maximum allowed number of clusters (2 out of 13), against 44% of the indices based on absolute or first min/max rule (24 out of 55).

Moreover, for complex data such as the considered metabolomics dataset, we verified that by doubling the maximum number of allowed clusters, that is, $K^{max} = 2 \cdot \sqrt{n}$, unreliable values of the optimal K are generally obtained due to possible degenerative behaviour of several indices, including those invariant to small K^{max} changes. Finally, when the number of variables is very high with respect to the number of samples, it is suggested to reduce the data dimensions because the correlation among them can significantly influence the similarity/diversity measures and, thus, the clustering outcomes.

CRediT authorship contribution statement

Roberto Todeschini: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. **Davide Ballabio:** Conceptualization, Validation, Visualization, Writing – review & editing. **Veronica Termopoli:** Conceptualization, Validation, Visualization, Writing – review & editing. **Viviana Consonni:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105117.

References

[1] L.F. Lago-Fernández, F. Corbacho, Normality-based validation for crisp clustering, Pattern Recogn. 43 (2010) 782–795.
[2] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J.M. Pérez, J.I. Martín, Towards a standard methodology to evaluate internal cluster validity indices, Pattern Recogn. Lett. 32 (2011) 505–515.
[3] G. Brock, V. Pihur, Su Datta, So Datta, clValid: An R Package for Cluster Validation, 2008.
[4] B. Desgraupes, Package ClusterCrit for R, 2013.
[5] E. Dimitriadou, K. Hornik, Package 'cclust', 2014.
[6] M. Walesiak, A. Dudek, Package 'clusterSim', 2014.
[7] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, NbClust: an R package for determining the relevant number of clusters in a data set, J. Stat. Software 61 (2014) 1–36.
[8] L. Nieweglowski, clv: Cluster Validation Techniques, 2020.
[9] G.W. Milligan, A Monte Carlo study of thirty internal criterion measures for cluster analysis, Psychometrika 46 (1981) 187–199.
[10] G.W. Milligan, M.C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (1985) 159–179.
[11] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2001) 107–145.
[12] M. Halkidi, M. Vazirgiannis, Clustering validity assessment: finding the optimal partitioning of a data set, in: Proceedings IEEE International Conference on Data Mining, 2001, pp. 187–194.
[13] S. Bandyopadhyay, M.K. Pakhira, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recogn. 37 (2004) 487–501.
[14] S. Bandyopadhyay, S. Saha, A point symmetry-based clustering technique for automatic evolution of clusters, IEEE Transactions on Knowledge and Data Engineering 20 (2008) 1441–1457.
[15] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recogn. 37 (2004) 487–501.
[16] M. Kim, R.S. Ramakrishna, New indices for cluster validity assessment, Pattern Recogn. Lett. 26 (2005) 2353–2363.
[17] Y. Tang, F. Sun, Improved validation index for fuzzy clustering, in: Proceedings of the American Control Conference, 2005, pp. 1121–1125.
[18] K.-L. Wu, M.-S. Yang, A cluster validity index for fuzzy clustering, Pattern Recogn. Lett. 26 (2005) 1275–1291.
[19] S. Saitta, B. Rapjaeel, I. Smith, A bounded index for cluster validity, in: P. Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, Springer, Berlin/Heidelberg (Germany), 2007.
[20] Y. Zhang, W. Wang, X. Zhang, Y. Li, A cluster validity index for fuzzy clustering, Inf. Sci. 178 (2008) 1205–1218.
[21] S. Saha, S. Bandyopadhyay, Performance evaluation of some symmetry-based cluster validity indexes, IEEE Trans. Syst. Man Cybern. C 39 (2009) 420–425.
[22] S. Sengupta, S. De, A. Konar, R. Janarthanan, An improved fuzzy clustering method using modifed Fukuyama-Sugeno cluster validity index, in: International Conference on Recent Trends in Information Systems, 2011, pp. 269–274.
[23] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perone, An extensive comparative study of cluster validity indices, Pattern Recogn. 46 (2013) 243–256.
[24] L.E. Brito da Silva, N.M. Melton, D.C. Wunsch II, Incremental cluster validity indices for hard partitions: extensions and comparative study, IEEE Access 8 (2020) 22025–22047.
[25] N. Wiroonsri, Clustering performance analysis using new correlation based cluster validity indices, 2021 arXiv:2109.11172v1.
[26] A.W.F. Edwards, L. Cavalli-Sforza, A method for cluster analysis, Biometrika 56 (1965) 362–375.
[27] G.H. Ball, D.J. Hall, Isodata, a Novel Method of Data Analysis and Pattern Classification, Menlo Park: Stanford Research Institute, NTIS No. AD 699616, 1965.
[28] J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (1993) 803–821.
[29] M.J. Symons, Clustering criteria and multivariate normal mixtures, Biometrics 37 (1981) 35–43.
[30] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1974) 1–27.
[31] J.A. Hartigan, Clustering Algorithms, Wiley, New York (NY, USA), 1975.
[32] D.A. Ratkowsky, G.N. Lance, A criterion for determining the number of groups in a classification, Aust. Comput. J. 10 (1978) 115–117.
[33] S.C. Sharma, Applied Multivariate Techniques, Wiley & Sons, New York (NY, USA), 1996.
[34] E. Zhu, X. Wang, F. Liu, A new cluster validity index for overlapping datasets, J. Phys. Conf. Ser. 1168 (2019) 032070.
[35] D.L. Davies, D.W. Bouldin, A clustering separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1979) 224–227.
[36] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, Proc. Fuzzy Syst. Symp. (1989) 247–250.
[37] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 841–846.
[38] S. Ray, R.H. Turi, Determination of number of clusters in k-means clustering and application in colour image segmentation, in: Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999, pp. 137–143.
[39] S.H. Kwon, Cluster validity index for fuzzy clustering, Electron. Lett. 34 (1998) 2176–2177.
[40] M.-S. Yang, K.-L. Wu, A new validity index for fuzzy clustering, 10th IEEE Int. Conf. Fuzzy Syst. 1 (2001) 89–92.
[41] K.R. Zalik, B. Zalik, Validity index for clusters of different sizes and densities, Pattern Recogn. Lett. 32 (2011) 221–234.
[42] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[43] B.R. Rezae, B.P.F. Lelieveldt, J. Reiber, A new cluster validity index for the fuzzy c-means, Pattern Recogn. Lett. 19 (1998) 237–246.
[44] D.-J. Kim, Y.-W. Park, D.-J. Park, A novel validity index for determination of the optimal number of clusters, IEEE Trans. Inf. Syst. E84-D (2001) 281–285.
[45] H.P. Friedman, J. Rubin, On some invariant criteria for grouping data, J. Am. Stat. Assoc. 62 (1967) 1159–1178.
[46] F.H.B. Marriot, Practical problems in a method of cluster analysis, Biometrics 27 (1975) 456–460.
[47] A.J. Scott, M.J. Symons, Clustering methods based on likelihood ratio criteria, Biometrics 27 (1971) 387–397.
[48] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 773–781.
[49] L.J. Hubert, J.R. Levin, A general statistic framework for assessing categorical clustering in free recall, Psychol. Bull. 83 (1976) 1072–1080.

[50] F.J. Rohlf, Methods of comparing classifications, Annu. Rev. Ecol. Syst. 5 (1974) 101–113.
[51] F.B. Baker, L.J. Hubert, Measuring the power of hierarchical cluster analysis, J. Am. Stat. Assoc. 70 (1975) 31–38.
[52] J.O. McClain, V.R. Rao, Clustisz: a program to test for the quality of clustering of a set of objects, J. Market. Res. 12 (1975) 456–460.
[53] I.J. Good, An index of separateness of clusters and a permutation test for its statistical significance, J. Stat. Comput. Simulat. 15 (1982) 81–84.
[54] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (1973) 32–57.
[55] J.C. Bezdek, N.R. Pal, Some new indexes of cluster validity, IEEE Trans. Syst. Man Cybern. B 28 (1998) 301–315.
[56] C.-H. Chou, M.-C. Su, E. Lai, A new cluster validity measure and its application to image compression, Pattern Anal. Appl. 7 (2004) 205–220.
[57] I. Gurrutxaga, I. Albisua, O. Arbelaitz, J.I. Martín, J. Muguerza, J.M. Pérez, I. Perona, SEP/COP: an efficient method to find the best partition in hierarchical clustering based on a new cluster validity index, Pattern Recogn. 43 (2010) 3364–3373.
[58] M. Kelly, R. Lomgjohn, K. Nottingham, The UCI Machine Learning Repository, 2023.
[59] H. Streuli, Mathematische Modelle für die chemische Zusammensetzung von Libensmitteln und ihre Bedeutung für deren Beurteilung, Lebensm. Technol. 20 (1987) 203–211.
[60] M. Forina, Personal Communication, 1990.
[61] A. Saviozzi, G. Lotti, D. Piacentini, La Composizione Amminoacidica Delle Farine Di Girasole, Rivista della Società Italiana di Scienze dell'Alimentazione, vol. 15, 1986, pp. 437–444.
[62] P.P. Mager, Design Statistics in Pharmacochemistry, Research Studies Press, Letchworth (UK), 1991.
[63] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188.
[64] M.J. Benito, M.C. Ortiz, M.S. Sánchez, L.A. Sarabia, M. Iñiguez, Typification of vinegars from Jerez and Rioja using classical chemometric techniques and neural network methods, Analyst 124 (1999) 547–552.
[65] M. Forina, C. Armanino, M. Castino, M. Ubigli, Multivariate data analysis as discriminating method of the origin of wines, Vitis 25 (1986) 189–201.
[66] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, Englewood Cliffs (NJ, USA), 1992.
[67] L. Kaufman, P.J. Rousseau. Finding Groups in Data. An Introduction to Cluster Analysis, Wiley & Sons, 1990.
[68] D. Brodnjak-Voncina, Z.C. Kodba, M. Novic, Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids, Chemom. Intell. Lab. Syst. 75 (2005) 31–43.
[69] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia, Classification of olive oils from their fatty acid composition, in: H. Martens, H. Russwurm Jr. (Eds.), Food Research and Data Analysis, Applied Science Publishers, London (UK), 1983.
[70] V. Batagelj, A. Mrvar, Pajek, 1996.
[71] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218.
[72] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (1971) 846–850.
[73] L.C. Morey, A. Agresti, The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement, Educ. Psychol. Meas. 44 (1984) 33–37.
[74] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[75] R. Bro, H.J. Nielsen, F. Savorani, K. Kjeldahl, I.J. Christensen, N. Brünner, A. J. Lawaetz, Data fusion in metabolomic cancer diagnostics, Metabolomics 9 (2013) 3–8.
[76] M. Buscema, G. Massini, M. Breda, W. Lodwick, F. Newman, M. Asadi-Zeydabadi, Artificial Adaptive Systems Using Auto Contractive Maps, Springer, Berlin (Ger), 2018.