Journal of Statistical Software

October 2014, Volume 61, Issue 6. *http://www.jstatsoft.org/*

NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set

Malika Charrad
Université de Gabes

Nadia Ghazzali
Université du Québec
à Trois-Rivières

Véronique Boiteau
Université Laval

Azam Niknafs
Université Laval

Abstract

Clustering is the partitioning of a set of objects into groups (clusters) so that objects within a group are more similar to each others than objects in different groups. Most of the clustering algorithms depend on some assumptions in order to define the subgroups present in a data set. As a consequence, the resulting clustering scheme requires some sort of evaluation as regards its validity.

The evaluation procedure has to tackle difficult problems such as the quality of clusters, the degree with which a clustering scheme fits a specific data set and the optimal number of clusters in a partitioning. In the literature, a wide variety of indices have been proposed to find the optimal number of clusters in a partitioning of a data set during the clustering process. However, for most of indices proposed in the literature, programs are unavailable to test these indices and compare them.

The R package **NbClust** has been developed for that purpose. It provides 30 indices which determine the number of clusters in a data set and it offers also the best clustering scheme from different results to the user. In addition, it provides a function to perform k-means and hierarchical clustering with different distance measures and aggregation methods. Any combination of validation indices and clustering methods can be requested in a single function call. This enables the user to simultaneously evaluate several clustering schemes while varying the number of clusters, to help determining the most appropriate number of clusters for the data set of interest.

Keywords: R package, cluster validity, number of clusters, clustering, indices, k-means, hierarchical clustering.

1. Introduction and related work

Clustering is the task of assigning a set of objects into groups (clusters) so that the objects in the same cluster are more similar to each other than objects in other clusters. There is a multitude of clustering methods available in the literature.

Everitt (1974) classified clustering methods into five basic types, namely: hierarchical methods, partitioning techniques, density or mode seeking techniques, clumping techniques and other methods not falling into the other categories. More recently, Sheikholeslami, Chatterjee, and Zhang (2000) classified clustering methods into the following types: partitional clustering, hierarchical clustering, density-based clustering and grid based clustering. Currently, there are several additional algorithms. Thus, we can distinguish *crisp* versus *fuzzy* clustering, *complete* versus *partial* clustering, *one-way* versus *two-way* clustering and *hierarchical* versus *partitional* clustering.

Most of the clustering algorithms take as input some parameters such as the number of clusters, the density of clusters or, at least, the number of points in a cluster.

Nonhierarchical procedures usually require the user to specify the number of clusters before any clustering is accomplished and hierarchical methods routinely produce a series of solutions ranging from n clusters to a solution with only one cluster present. As such, the problem of deciding on the number of clusters which suitably fit a data set, as well as the evaluation of the clustering results, have been subject to several research efforts. The procedure of evaluating the results of a clustering algorithm is known under the term *cluster validity.*

In Theodoridis and Koutroubas (2008), three approaches to investigate cluster validity are described. The first is based on external criteria, which consist in comparing the results of cluster analysis to externally known results, such as externally provided class labels. The second approach is based on internal criteria, which use the information obtained from within the clustering process to evaluate how well the results of cluster analysis fit the data without reference to external information. The third approach of clustering validity is based on relative criteria, which consists in the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameter values, e.g., the number of clusters.

A variety of measures aiming to validate the results of a clustering analysis have been defined and proposed in the literature for each of the approaches mentioned above. However, in this paper, we focus on indices proposed for the third approach.

Indeed, Milligan and Cooper (1985) examined thirty indices, with simulated data, where the number of clusters is known beforehand. Thirteen indices among them are available in R (R Core Team 2014) through the following packages: **cclust** (Dimitriadou 2014) and **clusterSim** (Walesiak and Dudek 2014).

In addition to indices described in the Milligan and Cooper (1985) study, Dunn (1974) introduced a validity index based on the distance between clusters and the diameter of the clusters and Rousseeuw and Kaufman proposed the "silhouette statistic" (Rousseeuw 1987; Kaufman and Rousseeuw 1990). More recently, Tibshirani, Walther, and Hastie (2001) proposed the "gap statistic". Lebart, Morineau, and Piron (2000) proposed a criterion based on the first and second derivatives and Halkidi, Vazirgiannis, and Batistakis (2000) and Halkidi and Vazirgiannis (2001) proposed two indices: SD index which is based on the concepts of average scattering for clusters and total separation between clusters (Halkidi *et al.* 2000), and SDbw index which is based on the criteria of compactness and separation between clusters (Halkidi and Vazirgiannis 2001).

However, as presented in Table 1, only nineteen indices among those mentioned above are implemented in the SAS cluster function (SAS Institute Inc. 2012) and in the following R pack-

	Index	SAS	**cclust**	**clusterSim**	**clv**	**clValid**
1	CH (Calinski and Harabasz 1974)	×		×		
2	CCC (Sarle 1983)	×				
3	Pseudot2 (Duda and Hart 1973)	×				
4	KL (Krzanowski and Lai 1988)			×		
5	Gamma (Baker and Hubert 1975)			×		
6	Gap (Tibshirani *et al.* 2001)			×		
7	Silhouette (Rousseeuw 1987)			×		
8	Hartigan (Hartigan 1975)		×	×		
9	Cindex (Hubert and Levin 1976)		×	×		
10	DB (Davies and Bouldin 1979)		×	×	×	
11	Ratkowsky (Ratkowsky and Lance 1978)		×			
12	Scott (Scott and Symons 1971)		×			
13	Marriot (Marriot 1971)		×			
14	Ball (Ball and Hall 1965)		×			
15	Trcovw (Milligan and Cooper 1985)		×			
16	Tracew (Milligan and Cooper 1985)		×			
17	Friedman (Friedman and Rubin 1967)		×			
18	Rubin (Friedman and Rubin 1967)		×			
19	Dunn (Dunn 1974)				×	×

Table 1: Indices implemented in SAS and R packages.

ages: **cclust** (Dimitriadou 2014), **clusterSim** (Walesiak and Dudek 2014), **clv** (Nieweglowski 2014) and **clValid** (Brock, Pihur, Datta, and Datta 2008; Brock, Pihur, and Datta 2014).

In this paper, we present a novel R package **NbClust**, which aims to gather all indices available in SAS or R packages together in only one package, and to include indices which are not implemented anywhere else in order to provide an exhaustive list of validity indices to estimate the number of clusters in a data set.

Some indices examined in the Milligan and Cooper (1985) study are not implemented in the **NbClust** package. Reasons for omission were that either not enough details were found to implement them or because they dependent on a certain method as it was considered desirable to examine only those indices that are method independent.

In the **NbClust** package, validity indices can be applied to outputs of two clustering algorithms: k-means and hierarchical agglomerative clustering (HAC), by varying all combinations of number of clusters, distance measures and clustering methods.

Distance measures available in **NbClust** package are: Euclidean distance, maximum distance, Manhattan distance, Canberra distance, binary distance and Minkowski distance. Several agglomeration methods are also provided by the **NbClust** package, namely: Ward (Ward 1963), single (Florek, Lukaszewicz, Perkal, and Zubrzycki 1951; Sokal and Michener 1958), complete (Sørensen 1948), average (Sokal and Michener 1958), McQuitty (McQuitty 1966), median (Gower 1967) and centroid (Sokal and Michener 1958). All of these methods and distance measures are described in detail in Section 3.

One important benefit of **NbClust** is that the user can simultaneously select multiple indices and number of clusters in a single function call. Moreover, it offers the user the best clustering scheme from different results. The package is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=NbClust (Charrad, Ghazzali, Boiteau, and Niknafs 2014).

The remainder of the paper is organized as follows. Section 2 provides a detailed description of validation measures available in **NbClust** package. Section 3 focuses on clustering algorithms implemented in **NbClust**. Section 4 gives an example of simulated and real data sets to illustrate the use of the **NbClust** package functions and objects. A brief conclusion follows in Section 5.

2. Clustering validity indices

Different clustering algorithms usually lead to different clusters of data; even for the same algorithm, the selection of different parameters or the presentation order of data objects may greatly affect the final clustering partitions. Thus, effective evaluation standards and criteria are critically important to give users confidence regarding the clustering results. At the same time, these assessments also provide some meaningful insights on how many clusters are hidden in the data.

In fact, in most real life clustering situations, the user faces the dilemma of selecting the number of clusters or partitions in the underlying data. As such, numerous indices for determining the number of clusters in a data set have been proposed.

All these clustering validity indices combine information about intracluster compactness and intercluster isolation, as well as other factors, such as geometric or statistical properties of the data, the number of data objects and dissimilarity or similarity measurements.

In the sequel, we present the indices implemented in the **NbClust** package and how to select the optimal number of clusters for each index.

In the following, we denote

$$
\begin{aligned}
n &= \text{ number of observations,}\\
p &= \text{ number of variables,}\\
q &= \text{ number of clusters,}\\
X &= \{x_{ij}\}\,,\ i = 1, 2, \ldots, n\,, j = 1, 2, \ldots, p,\\
&= n \times p \text{ data matrix of } p \text{ variables measured on } n \text{ independent observations,}\\
\overline{X} &= q \times p \text{ matrix of cluster means,}\\
\overline{x} &= \text{centroid of data matrix } X,\\
n_k &= \text{number of objects in cluster } C_k,\\
c_k &= \text{centroid of cluster } C_k,\\
x_i &= p\text{-dimensional vector of observations of the } i\text{th object in cluster } C_k,\\
\|x\| &= (x^\top x)^{1/2},
\end{aligned}
$$

W_q $= \sum_{k=1}^{q} \sum\limits_{i \in C_k} (x_i - c_k)(x_i - c_k)^\top$ is the within-group dispersion matrix for data clustered into q clusters,

B_q $= \sum_{k=1}^{q} n_k (c_k - \overline{x})(c_k - \overline{x})^\top$ is the between-group dispersion matrix for data clustered into q clusters,

N_t = total number of pairs of observations in the data set:

$$N_t = \frac{n(n-1)}{2},$$

N_w = total number of pairs of observations belonging to the same cluster:

$$N_w = \sum_{k=1}^{q} \frac{n_k(n_k-1)}{2},$$

N_b = total number of pairs of observations belonging to different clusters:

$$N_b = N_t - N_w,$$

S_w = sum of the within-cluster distances:

$$S_w = \sum_{k=1}^{q} \sum_{\substack{i,j \in C_k \\ i<j}} d(x_i, x_j),$$

S_b = sum of the between-cluster distances:

$$S_b = \sum_{k=1}^{q-1} \sum_{l=k+1}^{q} \sum_{\substack{i \in C_k \\ j \in C_l}} d(x_i, x_j).$$

2.1. CH index

The Calinski and Harabasz (CH) index (Calinski and Harabasz 1974) is defined by Equation 1.

$$\mathrm{CH}(q) = \frac{\mathrm{trace}(B_q)/(q-1)}{\mathrm{trace}(W_q)/(n-q)}. \tag{1}$$

The value of q, which maximizes $\mathrm{CH}(q)$, is regarded as specifying the number of clusters in Calinski and Harabasz (1974).

2.2. Duda index

Duda and Hart (1973) proposed a ratio criterion $Je(2)/Je(1)$ (Equation 2), where $Je(2)$ is the sum of squared errors within clusters when the data are partitioned into two clusters, and $Je(1)$ gives the squared errors when only one cluster is present.

$$\text{Duda} = \frac{Je(2)}{Je(1)} = \frac{W_k + W_l}{W_m}. \tag{2}$$

It is assumed that clusters C_k and C_l are merged to form C_m.

In Gordon (1999), the optimal number of clusters is the smallest q such that

$$\text{Duda} \geq 1 - \frac{2}{\pi p} - z\sqrt{\frac{2\left(1 - \frac{8}{\pi^2 p}\right)}{n_m p}} = \mathit{critValue\_Duda}, \tag{3}$$

where z is a standard normal score. Several values for the standard score were tested and the best results were obtained when the value was set to 3.20 (Milligan and Cooper 1985).

2.3. Pseudot2 index

Duda and Hart (1973) proposed another index, Pseudo t^2, which can only be applied to hierarchical methods. It is computed using Equation 4.

$$\text{Pseudot2} = \frac{V_{kl}}{\frac{W_k + W_l}{n_k + n_l - 2}}, \tag{4}$$

where $V_{kl} = W_m - W_k - W_l$, if $C_m = C_k \cup C_l$.

Gordon (1999) specified that the optimal number of clusters is the smallest q such that:

$$\text{Pseudot2} \leq \left(\frac{1 - \mathit{critValue\_Duda}}{\mathit{critValue\_Duda}}\right) \times (n_k + n_l - 2)\,. \tag{5}$$

2.4. Cindex

The C-Index was reviewed in Hubert and Levin (1976). It is calculated using Equation 6.

$$\text{Cindex} = \frac{S_w - S_{\min}}{S_{\max} - S_{\min}}, \; S_{\min} \neq S_{\max}, \; \text{Cindex} \in (0, 1), \tag{6}$$

where

- $S_{\min}$ = is the sum of the N_w smallest distances between all the pairs of points in the entire data set (there are N_t such pairs);
- $S_{\max}$ = is the sum of the N_w largest distances between all the pairs of points in the entire data set.

The minimum value of the index is used to indicate the optimal number of clusters (Milligan and Cooper 1985; Gordon 1999).

2.5. Gamma index

This index, calculated using Equation 7, represents an adaptation of Goodman and Kriskal's Gamma statistic for use in clustering situation (Baker and Hubert 1975).

Comparisons are made between all within-cluster dissimilarities and all between-cluster dissimilarities. A comparison is deemed to be concordant $[s(+)]$ (resp. discordant $[s(-)]$) if a within-cluster dissimilarity is strictly less (resp. strictly greater) than a between-cluster dissimilarity; equalities between members of two sets of dissimilarities are disregarded in the definition of the index (Gordon 1999).

$$\text{Gamma} = \frac{s(+) - s(-)}{s(+) + s(-)}, \tag{7}$$

where

- $s(+)$ = number of concordant comparisons,
- $s(-)$ = number of discordant comparisons.

The maximum value of the index is taken to represent the correct number of clusters (Milligan and Cooper 1985). In the **NbClust** package, this index is calculated only if the index argument is set to `"gamma"` or `"alllong"` because of its high computational demand.

2.6. Beale index

Beale (1969) proposed the use of an F-ratio to test the hypothesis of the existence of q_1 versus q_2 clusters in the data ($q_2 > q_1$).

Beale index is computed using Equation 8.

$$\text{Beale} = F \equiv \frac{\left(\frac{V_{kl}}{W_k+W_l}\right)}{\left(\left(\frac{n_m-1}{n_m-2}\right) 2^{\frac{2}{p}} - 1\right)}, \tag{8}$$

where $V_{kl} = W_m - W_k - W_l$. It is assumed that clusters C_k and C_l are merged to form C_m.

The optimal number of clusters is obtained by comparing F with an $F_{p,(n_m-2)p}$ distribution. The null hypothesis of a single cluster is rejected for significantly large values of F (Gordon 1999). By default, in our package, the 10% significance level was used to reject the null hypothesis (`alphaBeale = 0.1` in function `NbClust`).

2.7. CCC index

The Cubic Clustering Criterion (CCC) is the test statistic provided by the SAS software package (Sarle 1983). It is computed using Equation 9.

$$\text{CCC} = \ln\left[\frac{1 - \mathsf{E}\left(R^2\right)}{1 - R^2}\right] \frac{\sqrt{\frac{np^*}{2}}}{\left(0.001 + \mathsf{E}\left(R^2\right)\right)^{1.2}} \tag{9}$$

where

$$R^2 = 1 - \frac{\text{trace}(X^\top X - \overline{X}^\top Z^\top Z\overline{X})}{\text{trace}(X^\top X)}$$

- $X^\top X$ = total-sample sum-of-squares and crossproducts (SSCP) matrix ($p \times p$),
- $\overline{X} = (Z^\top Z)^{-1} Z^\top X$
- Z is a cluster indicator matrix ($n \times q$) with element $z_{ik} = 1$ if the ith observation belongs to the kth cluster and $z_{ik} = 0$ otherwise.

$$\mathsf{E}\left(R^2\right) = 1 - \left[\frac{\sum\limits_{j=1}^{p^*} \frac{1}{n+u_j} + \sum\limits_{j=p^*+1}^{p} \frac{u_j^2}{n+u_j}}{\sum\limits_{j=1}^{p} u_j^2}\right] \left[\frac{(n-q)^2}{n}\right] \left[1 + \frac{4}{n}\right].$$

- $u_j = \frac{s_j}{c}$,
- s_j = square root of the jth eigenvalue of $X^\top X/(n-1)$,
- $c = \left(\frac{v^*}{q}\right)^{\frac{1}{p^*}}$,
- $v^* = \prod\limits_{j=1}^{p^*} s_j$,
- p^* is chosen to be the largest integer less than q such that u_{p^*} is not less than one.

The maximum value of the index is used to indicate the optimal number of clusters in the data set (Milligan and Cooper 1985).

2.8. Ptbiserial index

This index, examined by Milligan (1980, 1981) and Kraemer (1982), is simply a point-biserial correlation between the raw input dissimilarity matrix and a corresponding matrix consisting of 0 or 1 entries. A value of 0 is assigned if the two corresponding points are clustered together by the algorithm. A value of one is assigned otherwise (Milligan 1980).

Given that larger positive values reflect a better fit between the data and the obtained partition, the maximum value of the index is used to select the optimal number of clusters in the data set (Milligan and Cooper 1985).

The point biserial correlation coefficient is calculated using Equation 10 (Milligan 1981).

$$\text{Ptbiserial} = \frac{\left[\overline{S}_b - \overline{S}_w\right]\left[N_w N_b / N_t^2\right]^{1/2}}{s_d}, \tag{10}$$

where

- $\overline{S}_w = S_w/N_w$,
- $\overline{S}_b = S_b/N_b$,
- s_d = standard deviation of all distances.

2.9. Gplus index

This index was reviewed by Rohlf (1974) and examined by Milligan (1981). It is computed using Equation 11.

$$\text{Gplus} = \frac{2s(-)}{N_t\left(N_t-1\right)}, \tag{11}$$

where $s(-)$ is the number of discordant comparisons, i.e., the number of times where two points which were in the same cluster had a larger distance than two points not clustered together (Milligan 1981). Minimum values of the index are used to determine the optimal number of clusters in the data (Milligan and Cooper 1985).

In the **NbClust** package, this index is calculated only if index argument is set to `"gplus"` or `"alllong"`, as it is computationally very expensive.

2.10. DB index

The Davies and Bouldin (1979) index is a function of the sum ratio of within-cluster scatter to between-cluster separation. It is calculated using Equation 12.

$$\text{DB}(q) = \frac{1}{q}\sum_{k=1}^{q} \max_{k \neq l}\left(\frac{\delta_k + \delta_l}{d_{kl}}\right), \tag{12}$$

where

- k, $l = 1, \ldots, q$ = cluster number,
- $d_{kl} = \sqrt[v]{\sum\limits_{j=1}^{p} \left|c_{kj} - c_{lj}\right|^v}$ = distance between centroids of clusters C_k and C_l (for $v = 2$, d_{kl} is the Euclidean distance),
- $\delta_k = \sqrt[u]{\frac{1}{n_k}\sum\limits_{i \in C_k}\sum\limits_{j=1}^{p} \left|x_{ij} - c_{kj}\right|^u}$ = dispersion measure of a cluster C_k (for $u = 2$, δ_k is the standard deviation of the distance of objects in cluster C_k to the centroid of this cluster).

The value of q minimizing $\text{DB}(q)$ is regarded as specifying the number of clusters (Milligan and Cooper 1985; Davies and Bouldin 1979).

2.11. Frey index

The index proposed by Frey and Van Groenewoud (1972), when they introduced their k-method of clustering, can only be applied to hierarchical methods. As shown in Equation 13, it is the ratio of difference scores from two successive levels in the hierarchy. The numerator is the difference between the mean between-cluster distances, $\overline{d}_b$, from each of the two hierarchy levels (level j and level $j+1$). The denominator is the difference between the mean within cluster distances, $\overline{d}_w$, from the two levels (level j and level $j+1$). The authors proposed, using a ratio score of 1.00, to identify the correct cluster level. The ratios often varied above and below 1.00.

The best results occurred when clustering was continued until the last ratio fell below 1.00. At this point, the cluster level before this was taken as optimal partition. If the ratio never fell below 1.00, a one cluster solution was assumed (Milligan and Cooper 1985).

$$\text{Frey} = \frac{\overline{S}_{b_{j+1}} - \overline{S}_{b_j}}{\overline{S}_{w_{j+1}} - \overline{S}_{w_j}}, \tag{13}$$

where

- $\overline{S}_b = S_b/N_b =$ mean between-cluster distance,
- $\overline{S}_w = S_w/N_w =$ mean within-cluster distance.

2.12. Hartigan index

The Hartigan index (Hartigan 1975) is computed using Equation 14.

$$\text{Hartigan} = \left(\frac{\text{trace}(W_q)}{\text{trace}(W_{q+1})} - 1\right)(n - q - 1), \tag{14}$$

where $q \in \{1, \ldots, n-2\}$. The maximum difference between hierarchy levels is taken as indicating the correct number of clusters in the data (Milligan and Cooper 1985).

2.13. Tau index

Tau index, reviewed by Rohlf (1974) and tested by Milligan (1981), is computed between corresponding entries in two matrices. The first contains the distances between items and the second 0/1 matrix indicates, whether or not, each pair of points are within the same cluster. Tau index is computed using Equation 15.

$$\text{Tau} = \frac{s(+) - s(-)}{\left[\left(N_t\left(N_t - 1\right)/2 - t\right)\left(N_t\left(N_t - 1\right)/2\right)\right]^{1/2}} \tag{15}$$

- $s(+)$ represents the number of times where two points not clustered together had a larger distance than two points which were in the same cluster, i.e., $s(+)$ is the number of concordant comparisons,
- $s(-)$ represents the reverse outcome (Milligan 1981), i.e., $s(-)$ is the number of discordant comparisons.
- N_t is the total number of distances and t is the number of comparisons of two pairs of points where both pairs represent within cluster comparisons or both pairs are between cluster comparisons.

The maximum value of the index is taken as indicating the correct number of clusters (Milligan and Cooper 1985). In the **NbClust** package, this index is calculated only if `index = "tau"` or `index = "alllong"`, because it is computationally very expensive.

2.14. Ratkowsky index

Ratkowsky and Lance (1978) proposed a criterion for determining the optimal number of clusters based on $\frac{\overline{S}}{q^{1/2}}$. The value of $\overline{S}$ is the average of the ratios of $(BGSS_j/TSS_j)$ where $BGSS$ stands for the sum of squares between the clusters (groups) for each variable and TSS for the total sum of squares for each variable (Hill 1980).

The optimal number of clusters is that value of q for which $\frac{\overline{S}}{q^{1/2}}$ has its maximum value (Milligan and Cooper 1985). If the value of q is made constant, the Ratkowsky and Lance criterion can be reduced from $\frac{\overline{S}}{q^{1/2}}$ to $\overline{S}$ (Hill 1980).

In the **NbClust** package, the Ratkowsky and Lance index is computed using Equation 16.

$$\text{Ratkowsky} = \frac{\overline{S}}{q^{1/2}}, \tag{16}$$

where

- $\overline{S}^2 = \frac{1}{p}\sum_{j=1}^{p} \frac{BGSS_j}{TSS_j}$,
- $BGSS_j = \sum_{k=1}^{q} n_k (c_{kj} - \overline{x}_j)^2$,
- $TSS_j = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$.

2.15. Scott index

Scott and Symons (1971) introduced an index based on Equation 17, where n is the number of elements in the data set, T is the total sum of squares and W_q is the sum of squares within the q clusters, as defined above.

$$\text{Scott} = n \log \frac{\det(T)}{\det(W_q)} \tag{17}$$

The maximum difference between hierarchy levels is used to suggest the correct number of partitions (Milligan and Cooper 1985).

2.16. Marriot index

Marriot (1971) proposed the following index calculated using Equation 18.

$$\text{Marriot} = q^2 \det(W_q). \tag{18}$$

The maximum difference between successive levels is used to determine the best partition level (Milligan and Cooper 1985).

2.17. Ball index

Ball and Hall (1965) proposed an index based on the average distance of the items to their respective cluster centroids. It is computed using Equation 19.

$$\text{Ball} = \frac{W_q}{q}, \tag{19}$$

(see also Dimitriadou, Dolnicar, and Weingessel 2002). The largest difference between levels is used to indicate the optimal solution (Milligan and Cooper 1985).

2.18. Trcovw index

This index, examined by Milligan and Cooper (1985), represents the trace of within clusters pooled covariance matrix. It is calculated using Equation 20.

$$\text{Trcovw} = \text{trace}\left(\mathsf{COV}\left(W_q\right)\right) \tag{20}$$

(see also Dimitriadou *et al.* 2002). Maximum difference scores between levels are used to indicate the optimal solution (Milligan and Cooper 1985).

2.19. Tracew index

This index has been one of the most popular indices suggested for use in clustering context (Milligan and Cooper 1985; Edwards and Cavalli-Sforza 1965; Friedman and Rubin 1967; Orloci 1967; Fukunaga and Koontz 1970). It is calculated using Equation 21:

$$\text{Tracew} = \text{trace}(W_q) \tag{21}$$

(see also Dimitriadou *et al.* 2002).

Given that the criterion increases monotonically with solutions containing fewer clusters, the maximum of the second differences scores are used to determine the number of clusters in the data (Milligan and Cooper 1985).

2.20. Friedman index

This index was proposed by Friedman and Rubin (1967), as a basis for a non hierarchical clustering method. It is computed using Equation 22.

$$\text{Friedman} = \text{trace}\left(W_q^{-1}B_q\right) \tag{22}$$

(see also Dimitriadou *et al.* 2002). The maximum difference in values of this criterion is used to indicate the optimal number of clusters (Milligan and Cooper 1985).

2.21. McClain index

The McClain and Rao index (McClain and Rao 1975) consists of the ratio of two terms (Equation 23). The first term is the average within cluster distance, divided by the number of within cluster distances. The denominator value is the average between cluster distance divided by the number of cluster distances.

$$\text{McClain} = \frac{\overline{S}_w}{\overline{S}_b} = \frac{S_w/N_w}{S_b/N_b}. \tag{23}$$

The minimum value of the index is used to indicate the optimal number of clusters.

2.22. Rubin index

Friedman and Rubin (1967) proposed another criterion based on the ratio of the determinant of the total sum of squares and cross products matrix to the determinant of the pooled within cluster matrix. This criterion is computed using Equation 24.

$$\text{Rubin} = \frac{\det(T)}{\det(W_q)} \tag{24}$$

(see also Dimitriadou *et al.* 2002). The minimum value of second differences between levels is used to select the optimal number of clusters (Milligan and Cooper 1985; Dimitriadou *et al.* 2002).

2.23. KL index

The KL index proposed by Krzanowski and Lai (1988) is defined by Equation 25.

$$\text{KL}(q) = \left| \frac{\text{DIFF}_q}{\text{DIFF}_{q+1}} \right|, \tag{25}$$

where $\text{DIFF}_q = (q-1)^{2/p}\, \text{trace}\,(W_{q-1}) - q^{2/p}\text{trace}\,(W_q)$. The value of q, maximizing $KL(q)$, is regarded as specifying the optimal number of clusters.

2.24. Silhouette index

Rousseeuw (1987) introduced the silhouette index computed using Equation 26.

$$\text{Silhouette} = \frac{\sum\limits_{i=1}^{n} S(i)}{n}, \; \text{Silhouette} \in [-1,\, 1]\,, \tag{26}$$

where

- $S(i) = \frac{b(i)-a(i)}{\max\{a(i);b(i)\}}$,
- $a(i) = \frac{\sum\limits_{j\in\{C_r \setminus i\}} d_{ij}}{n_r - 1}$ is the average dissimilarity of the ith object to all other objects of cluster C_r,
- $b(i) = \min\limits_{s \neq r} \{d_{iC_s}\}$,
- $d_{iC_s} = \frac{\sum\limits_{j\in C_s} d_{ij}}{n_s}$ is the average dissimilarity of the ith object to all objects of cluster C_s.

The maximum value of the index is used to determine the optimal number of clusters in the data (Kaufman and Rousseeuw 1990). $S(i)$ is not defined for $k = 1$ (only one cluster).

2.25. Gap index

The estimated Gap statistic proposed by Tibshirani *et al.* (2001) is computed using Equation 27.

$$\text{Gap}(q) = \frac{1}{B}\sum_{b=1}^{B} \log W_{qb} - \log W_q, \tag{27}$$

where B is the number of reference data sets generated using uniform prescription (Tibshirani *et al.* 2001) and W_{qb} is the within-dispersion matrix defined as in the Hartigan index. The optimal number of clusters is chosen via finding the smallest q such that:

$$\text{Gap}(q) \geq \text{Gap}(q+1) - s_{q+1}, \quad (q = 1, \ldots, n-2),$$

where

- $s_q = sd_q\sqrt{1 + 1/B}$,
- sd_q is the standard deviation of $\{\log W_{qb}\}$, $b = 1, \ldots, B$: $sd_q = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\log W_{qb} - \bar{l}\right)^2}$,
- $\bar{l} = \frac{1}{B}\sum_{b=1}^{B}\log W_{qb}$.

In the **NbClust** package, the Gap index is calculated only if `method = "gap"` or `method = "alllong"`, because of its high computational cost.

2.26. Dindex

The Dindex (Lebart *et al.* 2000) is based on clustering gain on intra-cluster inertia. Intra-cluster inertia measures the degree of homogeneity between the data associated with a cluster. It calculates their distances compared to the reference point representing the profile of the cluster, i.e., the cluster centroid in general. It can be defined using Equation 28.

$$w(P^q) = \frac{1}{q}\sum_{k=1}^{q}\frac{1}{n_k}\sum_{x_i \in C_k} d(x_i, c_k) \tag{28}$$

Given two partitions, P^{k-1} composed of $k-1$ clusters and P^k composed of k clusters, the clustering gain on intra-cluster inertia is defined as shown in Equation 29.

$$\text{Gain} = w(P^{q-1}) - w(P^q) \tag{29}$$

This clustering gain should be minimized.

The optimal cluster configuration can be identified by the sharp knee that corresponds to a significant decrease of the first differences of clustering gain versus the number of clusters. This knee or great jump of gain values can be identified by a significant peak in second differences of clustering gain.

2.27. Dunn index

The Dunn index (Dunn 1974) defines the ratio between the minimal intercluster distance to maximal intracluster distance. This index is given by Equation 30.

$$\text{Dunn} = \frac{\min\limits_{1 \leq i < j \leq q} d(C_i, C_j)}{\max\limits_{1 \leq k \leq q} \text{diam}(C_k)}, \tag{30}$$

where $d(C_i, C_j)$ is the dissimilarity function between two clusters C_i and C_j defined as $d(C_i, C_j) = \min\limits_{x \in C_i, y \in C_j} d(x, y)$ and $\text{diam}(C)$ is the diameter of a cluster, which may be considered as a measure of cluster dispersion. The diameter of a cluster C can be defined using Equation 31.

$$\text{diam}(C) = \max_{x,y \in C} d(x, y) \tag{31}$$

If the data set contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large. Thus, Dunn index should be maximized.

2.28. Hubert statistic

Hubert's Γ statistic (Hubert and Arabie 1985) is the point serial correlation coefficient between any two matrices. When the two matrices are symmetric, Γ can be written in its raw form as shown by Equation 32.

$$\Gamma(P, Q) = \frac{1}{N_t} \sum_{\substack{i=1 \\ i<j}}^{n-1} P_{ij} Q_{ij}, \tag{32}$$

where

- P is the proximity matrix of the data set,
- Q is an $n \times n$ matrix whose (i, j) element is equal to the distance between the representative points (v_{c_i}, v_{c_j}) of the clusters where the objects x_i and x_j belong.

We note that for $q = 1$ or $q = n$, the index is not defined.

The definition of Hubert's normalized Γ statistic is given by Equation 33.

$$\overline{\Gamma} = \frac{\sum\limits_{\substack{i=1 \\ i<j}}^{n-1} (P_{ij} - \mu_P)(Q_{ij} - \mu_Q)}{\sigma_P \sigma_Q}, \tag{33}$$

where μ_P, μ_Q, σ_P, σ_Q are the respective means and variances of the P and Q matrices.

This index takes values between -1 and 1. If P and Q are not symmetric then all summations are extended over all n^2 entries and $N_t = n^2$ (Bezdek and Pal 1998).

High values of normalized Γ statistics indicate the existence of compact clusters. Thus, in the plot of normalized Γ versus q (q is the number of clusters), we seek a significant knee that corresponds to a significant increase of normalized Γ as q varies from 2 to $q_{\max}$, where $q_{\max}$ is the maximum possible number of clusters. The number of clusters at which the knee occurs is an indication of the number of clusters that underlie the data (Halkidi, Batistakis, and Vazirgiannis 2001).

In the **NbClust** package, second differences values of normalized Γ statistics are plotted to help distinguish the knee from other anomalies. A significant peak in this plot indicates the optimal number of clusters.

2.29. SDindex

The SD validity index definition is based on the concepts of *average scattering for clusters* and *total separation between clusters*. It is computed using Equation 34.

$$\text{SDindex}(q) = \alpha \text{Scat}(q) + \text{Dis}(q) \tag{34}$$

The first term, Scat(q), calculated using Equation 35, indicates the average compactness of clusters (i.e., intra-cluster distance). A small value for this term indicates compact clusters.

$$\text{Scat}(q) = \frac{\frac{1}{q}\sum_{k=1}^{q} \left\| \sigma^{(k)} \right\|}{\|\sigma\|} \tag{35}$$

where

- σ is the vector of variances for each variable in the data set, $\sigma = (\mathsf{VAR}(V_1), \mathsf{VAR}(V_2), \ldots, \mathsf{VAR}(V_p))$,
- $\sigma^{(k)}$ is the variance vector for each cluster C_k, $\sigma^{(k)} = (\mathsf{VAR}(V_1^{(k)}), \mathsf{VAR}(V_2^{(k)}), \ldots, \mathsf{VAR}(V_p^{(k)}))$.

The second term Dis(q), calculated using Equation 36, indicates the total separation between the q clusters (i.e., an indication of inter-cluster distance).

$$\text{Dis}(q) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^{q} \left(\sum_{z=1}^{q} \|c_k - c_z\| \right)^{-1} \tag{36}$$

where

- $D_{\max} = \max\left(\|c_k - c_z\|\right) \forall k, z \in \{1, 2, 3, \ldots, q\}$ is the maximum distance between cluster centers,
- $D_{\min} = \min\left(\|c_k - c_z\|\right) \forall k, z \in \{1, 2, 3, \ldots, q\}$ is the minimum distance between cluster centers.

α is a weighting factor equal to Dis$(q_{\max})$ where $q_{\max}$ is the maximum number of input clusters. The number of clusters, q, that minimizes the above index, can be considered as an optimal value for the number of clusters present in the data set.

In the **clv** package, SD index is programmed with α equal to $q_{\max}$. Conversely, in the **NbClust** package, α is equal to $\text{Dis}(q_{\max})$ as mentioned in Halkidi *et al.* (2000).

2.30. SDbw index

The SDbw validity index definition is based on the criteria of compactness and separation between clusters. It is computed using Equation 37.

$$\text{SDbw}(q) = \text{Scat}(q) + \text{Density.bw}(q) \tag{37}$$

The first term, $\text{Scat}(q)$, is the same computed in SDindex (Equation 34).

The second term, $\text{Density.bw}(q)$, is the inter-cluster density. It evaluates the average density in the region among clusters in relation to the density of the clusters and it is calculated using Equation 38.

$$\text{Density.bw}(q) = \frac{1}{q(q-1)} \sum_{i=1}^{q} \left(\sum_{j=1, i \neq j}^{q} \frac{\text{density}(u_{ij})}{\max(\text{density}(c_i), \text{density}(c_j))} \right), \tag{38}$$

where

- u_{ij} is the middle point of the line segment defined by the clusters centroids c_i and c_j,
- $\text{density}(u_{ij})$ is calculated using Equation 39.

$$\text{density}(u_{ij}) = \sum_{l=1}^{n_{ij}} f(x_l, u_{ij}), \tag{39}$$

where

- n_{ij} is the number of tuples that belong to the clusters C_i and C_j,
- $f(x_l, u_{ij})$ is equal to 0 if $d(x, u_{ij}) > \text{Stdev}$ and 1 otherwise,
- Stdev, defined in Equation 40, is the average standard deviation of clusters.

$$\text{Stdev} = \frac{1}{q} \sqrt{\sum_{k=1}^{q} \left\| \sigma^{(k)} \right\|} \tag{40}$$

The number of clusters q that minimizes SDbw is considered as the optimal value for the number of clusters in the data set (Halkidi and Vazirgiannis 2001).

As mentioned above, the optimal number of clusters selected by **NbClust** for each index is based on maximum (or minimum) values of the index, maximum (or minimum) difference between hierarchy levels of the index ($\max_q(i_q - i_{q-1})$, q is the number of clusters and i_q is the index value for q clusters), maximum (or minimum) value of second differences between levels of the index ($\max_q((i_{q+1} - i_q) - (i_q - i_{q-1}))$) or by the use of a critical value such as in the case of the Gap index and the Beale index.

If the measure increases as the number of clusters increases, such as in the case of the Dindex and the Hubert index, then simply finding the minimum or maximum on a plot is no longer sufficient. Instead, a significant local change in the value of the measure, seen as a "knee" in the plot, indicates the best parameters for clustering.

In the **NbClust** package, the knee is detected by a local peak in the plot of second differences between levels of the index. Thus, the suitable number of clusters is chosen by visual inspection of the second differences plot. The absence of such a knee might be an indication that the data set possesses no clustering structure.

Table 2 summarizes the indices included in **NbClust** package. It gives the name of each index in references and in the **NbClust** package, and how to select the optimal number of clusters.

3. Clustering algorithms

There is a multitude of clustering methods available in the literature which can be classified into different types (see also Section 1). For each of the types there are various of subtypes and different algorithms for finding clusters in a data set (Jain, Murty, and Flyn 1998; Halkidi *et al.* 2000; Theodoridis and Koutroubas 2008). The R project for statistical computing provides a wide variety of these clustering algorithms either through the base distribution or add-on packages.

Currently, k-means (MacQueen 1967; Hartigan and Wong 1979) is one of the most popularly adopted partitioning algorithms, as evidenced by its use in a wide variety of packages in the R system for statistical computing, such as **cclust** (Dimitriadou 2014), **clustTool** (Templ 2007), **clue** (Hornik 2005, 2014), among others. An alternative approach for partitioning clustering is hierarchical clustering, which is a widely used clustering method, as seen in many R packages such as **hybridHclust** (Chipman and Tibshirani 2014), **pvclust** (Suzuki and Shimodaira 2014) and **cluster** (Maechler, Rousseeuw, Struyf, and Hubert 2014).

In the current version of the **NbClust** package, only k-means and the agglomerative approach of hierarchical clustering are available. Next versions will include other clustering methods such as self organizing maps.

In the following, a brief description of k-means and hierarchical agglomerative clustering algorithms is provided.

3.1. k-means

k-means is an iterative method which minimizes the within-class sum of squares for a given number of clusters (MacQueen 1967; Hartigan and Wong 1979). The algorithm starts with an initial guess for cluster centers, and each observation is placed in the cluster to which it is closest. The cluster centers are then updated, and the entire process is repeated until the cluster centers no longer move. Often another clustering algorithm (e.g., UPGMA) is run initially to determine starting points for the cluster centers. k-means is said to be a reallocation method. Here is the general principle:

1. Select as many points as the number of desired clusters to create initial centers.

2. Each observation is then associated with the nearest center to create temporary clusters.

3. The gravity centers of each temporary cluster are calculated and these become the new cluster centers.

	Name of the index in **NbClust**	Optimal number of clusters
1.	`"ch"` (Calinski and Harabasz 1974)	Maximum value of the index
2.	`"duda"` (Duda and Hart 1973)	Smallest number of clusters such that index > criticalValue
3.	`"pseudot2"` (Duda and Hart 1973)	Smallest number of clusters such that index < criticalValue
4.	`"cindex"` (Hubert and Levin 1976)	Minimum value of the index
5.	`"gamma"` (Baker and Hubert 1975)	Maximum value of the index
6.	`"beale"` (Beale 1969)	Number of clusters such that critical value >= alpha
7.	`"ccc"` (Sarle 1983)	Maximum value of the index
8.	`"ptbiserial"` (Milligan 1980, 1981)	Maximum value of the index
9.	`"gplus"` (Rohlf 1974; Milligan 1981)	Minimum value of the index
10.	`"db"` (Davies and Bouldin 1979)	Minimum value of the index
11.	`"frey"` (Frey and Van Groenewoud 1972)	Cluster level before index value < 1.00
12.	`"hartigan"` (Hartigan 1975)	Maximum difference between hierarchy levels of the index
13.	`"tau"` (Rohlf 1974; Milligan 1981)	Maximum value of the index
14.	`"ratkowsky"` (Ratkowsky and Lance 1978)	Maximum value of the index
15.	`"scott"` (Scott and Symons 1971)	Maximum difference between hierarchy levels of the index
16.	`"marriot"` (Marriot 1971)	Max. value of second differences between levels of the index
17.	`"ball"` (Ball and Hall 1965)	Maximum difference between hierarchy levels of the index
18.	`"trcovw"` (Milligan and Cooper 1985)	Maximum difference between hierarchy levels of the index
19.	`"tracew"` (Milligan and Cooper 1985)	Max. value of second differences between levels
20.	`"friedman"` (Friedman and Rubin 1967)	Maximum difference between hierarchy levels of the index
21.	`"mcclain"` (McClain and Rao 1975)	Minimum value of the index
22.	`"rubin"` (Friedman and Rubin 1967)	Minimum value of second differences between levels
23.	`"kl"` (Krzanowski and Lai 1988)	Maximum value of the index
24.	`"silhouette"` (Rousseeuw 1987)	Maximum value of the index
25.	`"gap"` (Tibshirani *et al.* 2001)	Smallest number of clusters such that criticalValue >= 0
26.	`"dindex"` (Lebart *et al.* 2000)	Graphical method
27.	`"dunn"` (Dunn 1974)	Maximum value of the index
28.	`"hubert"` (Hubert and Arabie 1985)	Graphical method
29.	`"sdindex"` (Halkidi *et al.* 2000)	Minimum value of the index
30.	`"sdbw"` (Halkidi and Vazirgiannis 2001)	Minimum value of the index

Table 2: Overview of the indices implemented in the **NbClust** package.

4. Each observation is reallocated to the cluster which has the closest center.
5. This procedure is iterated until convergence.

3.2. Hierarchical clustering

Hierarchical clustering seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative or "bottom up" approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive or "top down" approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

Hierarchical clustering requires to define a dissimilarity measure (or distance) and an agglomeration criterion. Many distances are available (Manhattan, Euclidean, etc.) as well as several agglomeration methods (Ward, single, centroid, etc.).

Dissimilarity measures

The following distance measures are written for two vectors x and y and are used when the data is a d-dimensional vector arising from measuring d characteristics on each of n objects or individuals (Seber 1984). The characteristics or variables may be quantitative (discrete or continuous) or qualitative (ordinal or nominal) (Seber 1984).

- Euclidean distance: it is the usual square distance between the two vectors. It is given by Equation 41.

$$d(x,y) = \left(\sum_{j=1}^{d} (x_j - y_j)^2 \right)^{\frac{1}{2}} \tag{41}$$

- Maximum distance: it is the maximum distance between two components of x and y (supremum norm), as described by Equation 42.

$$d(x,y) = \sup_{1 \le j \le d} |x_j - y_j| \tag{42}$$

- Manhattan distance: is the absolute distance between the two vectors. It is given by Equation 43.

$$d(x,y) = \sum_{j=1}^{d} |x_j - y_j| \tag{43}$$

- Canberra distance: terms with zero numerator and denominator are omitted from the sum and treated as if the values were missing.

$$d(x,y) = \sum_{j=1}^{d} \frac{|x_j - y_j|}{|x_j| + |y_j|} \tag{44}$$

- Binary distance: the vectors are regarded as binary bits, so non-zero elements are “on” and zero elements are “off”. The distance is the proportion of bits in which only one is on amongst those in which at least one is on.
- Minkowski distance: is the p norm, i.e., the pth root of the sum of the pth powers of the differences of the components.

$$d(x,y) = \left(\sum_{j=1}^{d} |x_j - y_j|^p \right)^{\frac{1}{p}} \tag{45}$$

Agglomeration methods

Most hierarchical clustering algorithms are variants of the single-link, complete-link, and minimum-variance algorithms. The following aggregation methods are available in **NbClust**.

- Ward (Ward 1963): Ward’s method minimizes the total within-cluster variance. At each step, the pair of clusters with minimum cluster distance is merged. This pair of clusters leads to minimum increase in total within-cluster variance after merging.

  Two algorithms, Ward1 and Ward2, are found in the literature and are available in software packages, both claiming that they implement the Ward clustering method. However, when applied to the same distance matrix D, they produce different results (Murtagh and Legendre 2011).

  The one used by option `"ward.D"`, equivalent to the only Ward option `"ward"` in R versions $\leq$ 3.0.3, does not implement Ward’s (1963) clustering criterion, whereas option `"ward.D2"` implements that criterion (Murtagh and Legendre 2014). With the latter, the dissimilarities are squared before cluster updating.

- Single (Florek *et al.* 1951; Sokal and Michener 1958): the distance D_{ij} between two clusters C_i and C_j is the minimum distance between two points x and y, with $x \in C_i$ and $y \in C_j$:

$$D_{ij} = \min_{x \in C_i,\, y \in C_j} d(x,y). \tag{46}$$

  A drawback of this method is the so-called chaining phenomenon: clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other. Consequently, this method often creates irregular and very elongated clusters.

- Complete (Sørensen 1948): the distance D_{ij} between two clusters C_i and C_j is the maximum distance between two points x and y, with $x \in C_i$ and $y \in C_j$:

$$D_{ij} = \max_{x \in C_i,\, y \in C_j} d(x,y). \tag{47}$$

- Average (Sokal and Michener 1958): the distance D_{ij} between two clusters C_i and C_j is the mean of the distances between the pair of points x and y, where $x \in C_i$ and $y \in C_j$:

$$D_{ij} = \sum_{x \in C_i,\, y \in C_j} \frac{d(x,y)}{n_i \times n_j}, \tag{48}$$

where n_i and n_j are respectively the number of elements in clusters C_i and C_j.

This method has the tendency to form clusters with the same variance and, in particular, small variance.

- McQuitty (McQuitty 1966): the distance between clusters C_i and C_j is the weighted mean of the between-cluster dissimilarities:

$$D_{ij} = \left(D_{ik} + D_{il}\right)/2, \tag{49}$$

where cluster C_j is formed from the aggregation of clusters C_k and C_l.

- Median (Gower 1967): the distance D_{ij} between two clusters C_i and C_j is given by the following formula:

$$D_{ij} = \frac{(D_{ik} + D_{il})}{2} - \frac{D_{kl}}{4}, \tag{50}$$

where cluster C_j is formed by the aggregation of clusters C_k and C_l.

- Centroid (Sokal and Michener 1958): the distance D_{ij} between two clusters C_i and C_j is the squared Euclidean distance between the gravity centers of the two clusters, i.e., between the mean vectors of the two clusters, $\bar{x}_i$ and $\bar{x}_j$ respectively:

$$D_{ij} = \left\|\bar{x}_i - \bar{x}_j\right\|^2. \tag{51}$$

This method is more robust than others in terms of isolated points.

4. Finding the relevant number of clusters using NbClust

In this section, we use a simulated and a real data set to show how the **NbClust** package works.

4.1. Simulated data set

We consider a simulated data set composed of 4 distinct nonoverlapping clusters (Figure 1). The data set consists of 200 points and the clusters are embedded in a bidimensional Euclidean space.

In R, a typical call for using **NbClust** is:

```
R> library("NbClust")
R> NbClust(data, diss = NULL, distance = "euclidean", min.nc = 2, max.nc = 8,
+    method = "complete", index = "alllong", alphaBeale = 0.1)
```

The function documentation regarding explicit instruction on input arguments is given online by the command `help(NbClust)`.

Our goal is to cluster rows of the data matrix based on columns (variables) and to evaluate the ability of available indices to identify the optimal number of clusters in the underlying data.

The number of clusters varies from 2 to 8. The distance metric (both for the applicable clustering methods and validation measures) is set to `"euclidean"`; other available options

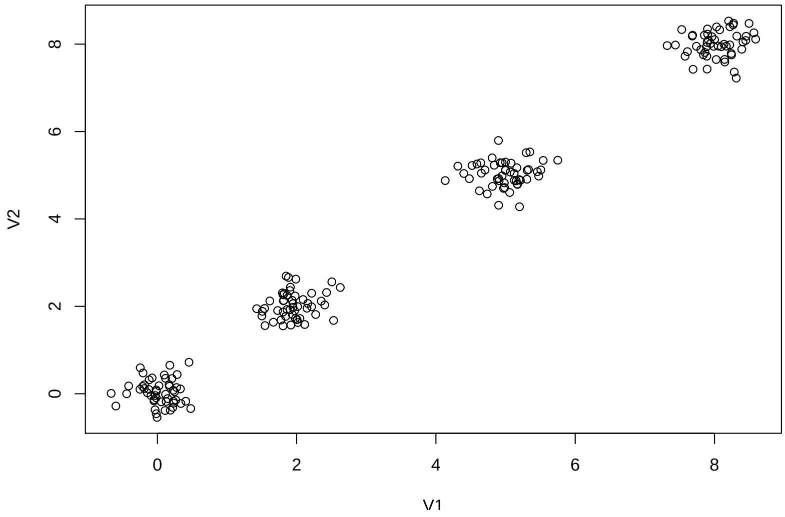

Figure 1: Simulated data set plot.

are "maximum", "manhattan", "camberra", "binary" and "minkowski". The agglomeration method for hierarchical clustering is set to "ward.D2". It is also possible to select another method such as "ward.D", "complete", "single", "mcquitty", "average", "median" or "centroid".

User can request indices one by one, by setting the argument index to the name of the index as presented in Table 2, for example index = "duda". In this case, as shown in the example below, **NbClust** function displays the *Duda* values of the partitions obtained with number of clusters ranging from min.nc to max.nc (from 2 to 8 in this example), the critical value of the *Duda* index for each partition, the best number of clusters, given in this case by the smallest number of clusters such that index > critical value (4 clusters in this example) and the partition corresponding to the best number of clusters.

```
R> library("NbClust")
R> res <- NbClust(data, distance = "euclidean", min.nc = 2, max.nc = 8,
+    method = "ward.D2", index = "duda")
R> res$All.index

All 200 observations were used.
$All.index
     2      3      4      5      6      7      8
0.0388 0.0738 0.5971 0.6691 0.6602 0.6210 0.4200

R> res$All.CriticalValues

$All.CriticalValues
     2      3      4      5      6      7      8
0.4349 0.4349 0.3327 0.3327 0.3327 0.3327 0.2234

R> res$Best.nc
```

```
$Best.nc
Number_clusters     Value_Index
         4.0000          0.5971

R> res$Best.partition

$Best.partition
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
[106] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[141] 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[176] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
```

Clustering with index argument set to "alllong" requires more time, as the run of some measures, such as Gamma, Tau, Gap and Gplus, is computationally very expensive, especially when the number of clusters and objects in the data set grows very large. The user can avoid running these four indices by setting the argument index to "all". In this case, only 26 indices are computed.

With the "alllong" option, the output of the NbClust function consists of

- *all* validation measures,
- critical values for Duda, Gap, PseudoT2 and Beale indices,
- the number of clusters corresponding to the optimal score for each measure,
- the best number of clusters proposed by NbClust according to the majority rule,
- the best partition.

```
R> NbClust(data, distance = "euclidean", min.nc = 2, max.nc = 8,
+    method = "complete", index = "alllong")

*** : The Hubert index is a graphical method of determining the number of
      clusters.
      In the plot of Hubert index, we seek a significant knee that
      corresponds to a significant increase of the value of the measure i.e
      the significant peak in Hubert index second differences plot.

*** : The D index is a graphical method of determining the number of
      clusters.
      In the plot of D index, we seek a significant knee (the significant
      peak in Dindex second differences plot) that corresponds to a
      significant increase of the value of the measure.

All 200 observations were used.
```

```
*******************************************************************
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 19 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4


*******************************************************************
$All.index
       KL        CH  Hartigan     CCC     Scott   Marriot   TrCovW
2  3.4594  885.3975  398.1129 17.3825  530.5325 39798.188 228.9249
3  1.2566 1524.1400 1126.0955 17.4487  761.9166 28157.985 237.8424
4 26.9498 7163.1259   17.0799 30.6642 1247.1374  4424.207 223.0371
5  0.9651 5814.9538   11.1679 26.0102 1287.0337  5662.677 151.7204
6  2.7236 4895.3856   22.8628 22.4624 1306.3867  7402.183 145.2914
7  2.1804 4540.5372   23.4182 20.3053 1351.0258  8059.734 123.0103
8  1.2075 4344.8386   20.6834 18.7214 1403.6219  8092.706 106.9838
    TraceW Friedman    Rubin Cindex     DB Silhouette   Duda  Pseudot2
2 676.8594 295.4549  13.8329 0.3601 0.4635     0.7209 0.0388 2430.0879
3 224.8201 328.3267  41.6464 0.3043 0.2878     0.7794 0.0738 1230.0792
4  33.4742 564.6962 279.7063 0.3058 0.2415     0.8441 0.7175   18.9003
5  30.7910 634.2843 304.0806 0.3428 0.7714     0.6948 0.7679   14.5044
6  29.1231 661.0045 321.4956 0.3515 0.9099     0.5547 0.6189   29.5544
7  26.0528 739.4890 359.3836 0.3425 1.0936     0.4442 0.6772   22.8848
8  23.2337 858.2385 402.9905 0.3238 1.2702     0.2986 0.4896   28.1467
    Beale Ratkowsky     Ball Ptbiserial    Gap    Frey McClain  Gamma
2 24.5463    0.6392 338.4297     0.8002 0.0643  0.9342  0.2645 0.9375
3 12.4250    0.5595  74.9400     0.7801 0.3547  0.9331  0.3067 0.9980
4  0.3857    0.4977   8.3685     0.7016 1.7257 12.9402  0.2496 1.0000
5  0.2960    0.4453   6.1582     0.6461 1.3799 10.3583  0.2972 0.9722
6  0.6032    0.4066   4.8538     0.6219 0.9882  8.5647  0.3211 0.9620
7  0.4670    0.3766   3.7218     0.5672 0.8816  5.1583  0.3845 0.9423
8  1.0052    0.3524   2.9042     0.5140 0.6793  4.3971  0.4548 0.9320
     Gplus      Tau   Dunn Hubert SDindex Dindex   SDbw
2 155.3597 4664.155 0.5009  3e-04  1.2374 1.7764 0.1828
3   4.7469 4638.747 0.6723  3e-04  0.7843 0.8928 0.0438
4   0.0011 3693.465 0.8184  4e-04  0.9362 0.3622 0.0091
5  46.8435 3272.053 0.0934  4e-04  5.9589 0.3455 0.0915
6  61.0775 3089.934 0.0975  4e-04  5.6107 0.3344 0.0895
7  81.9910 2680.056 0.0628  4e-04  6.0590 0.3152 0.1373
8  83.6208 2293.822 0.0640  4e-04  5.3941 0.2994 0.1280
```

```
$All.CriticalValues
  CritValue_Duda CritValue_PseudoT2 Fvalue_Beale CritValue_Gap
2         0.4349           127.3323       0.0000       -0.2859
3         0.4349           127.3323       0.0000       -1.3651
4         0.3327            96.2763       0.6810        0.3531
5         0.3327            96.2763       0.7445        0.4008
6         0.3327            96.2763       0.5491        0.1166
7         0.3327            96.2763       0.6283        0.2139
8         0.2234            93.8395       0.3727        0.1015
```

Critical values are used to select the best number of clusters. For example, the optimal number of clusters proposed by the *Duda* index is the smallest number of clusters such that critical value is less than the index value (e.g., 4 clusters in the example above).

```
$Best.nc
                    KL       CH Hartigan     CCC    Scott  Marriot
Number_clusters  4.0000    4.000    4.000  4.0000   4.0000     4.00
Value_Index     26.9498 7163.126 1109.016 30.6642 485.2208 24972.25
                 TrCovW   TraceW Friedman     Rubin Cindex     DB
Number_clusters  5.0000   3.0000   4.0000    4.0000 3.0000 4.0000
Value_Index     71.3167 260.6934 236.3695 -213.6856 0.3043 0.2415
                Silhouette   Duda PseudoT2  Beale Ratkowsky     Ball
Number_clusters     4.0000 4.0000   4.0000 4.0000    2.0000   3.0000
Value_Index         0.8441 0.7175  18.9003 0.3857    0.6392 263.4897
                PtBiserial    Gap Frey McClain Gamma  Gplus      Tau
Number_clusters     2.0000 4.0000    1  4.0000     4 4.0000    2.000
Value_Index         0.8002 1.7257   NA  0.2496     1 0.0011 4664.155
                  Dunn Hubert SDindex Dindex   SDbw
Number_clusters 4.0000      0  3.0000      0 4.0000
Value_Index     0.8184      0  0.7843      0 0.0091


$Best.partition
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
[106] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[141] 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[176] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
```

As mentioned in Section 2, the Dindex and the Hubert index are graphical methods. Hence, values of these indices are set to zero in the example above. In this case, the optimal number of clusters is identified by a significant knee in the plot of index values against number of clusters. This knee corresponds to a significant increase or significant decrease of the index, as the number of clusters varies from the minimum to the maximum. In the **NbClust** package,

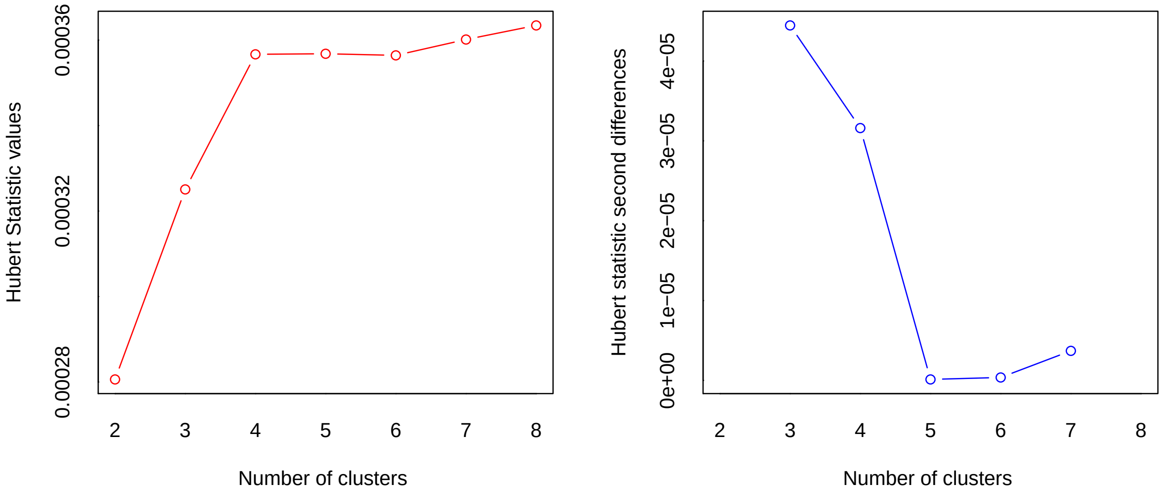

Figure 2: Hubert statistic graphic for determining the best number of clusters in the simulated data set.

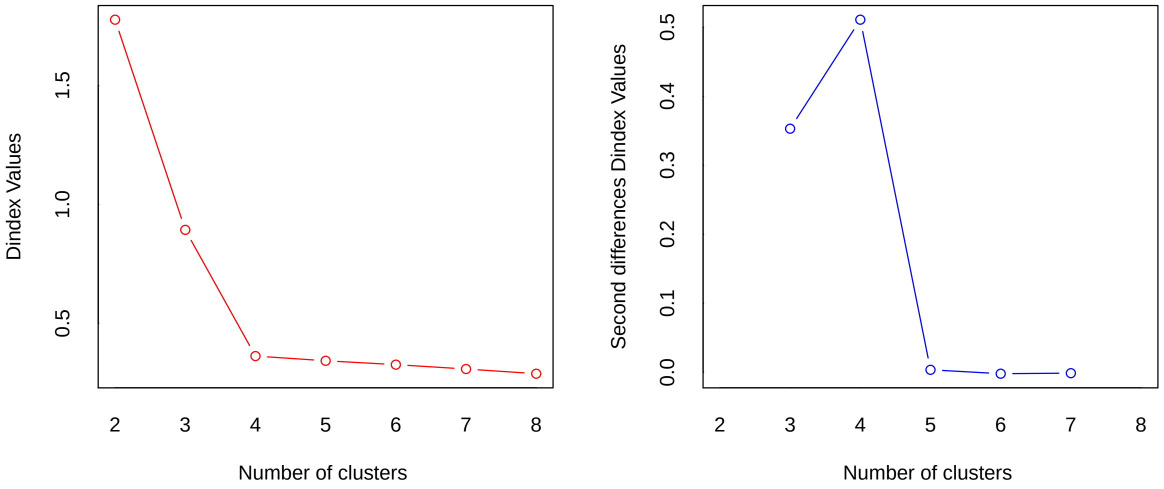

Figure 3: Dindex graphic for determining the best number of clusters in the simulated data set.

a significant peak in the plot of second differences values indicates the relevant number of clusters.

As shown in Figures 2 and 3, the Hubert index proposes 3 as the best number of clusters and the Dindex proposes 4 as the best number of clusters.

Certainly, the results presented in the example above seem to indicate that there is no unanimous choice regarding the optimal number of clusters. Indeed, 20 among 30 indices propose 4 as the best number of clusters, 5 indices propose 3 as the optimal number of clusters, 3 indices select 2 as the relevant number of clusters in this data set and only one index proposes 5 as the best number of clusters. Consequently, the user faces the dilemma of choosing one

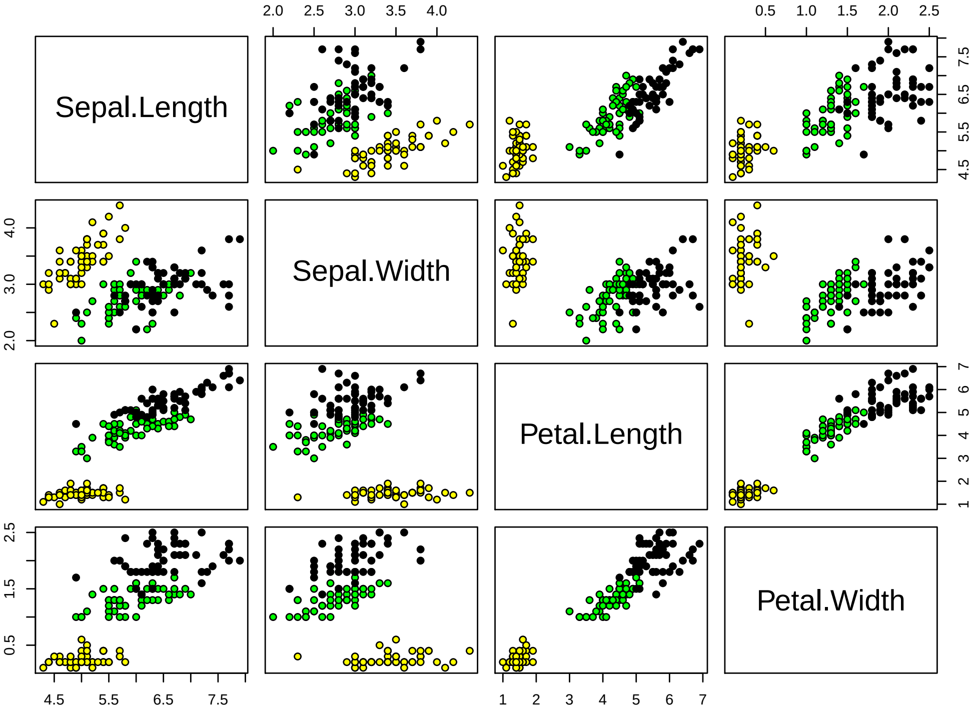

Figure 4: Pairwise scatter plots for the original Iris data.

among four available solutions (2, 3, 4 or 5 clusters).

There are two ways to deal with this problem. The first one is based on the majority rule, which is available in the **NbClust** package. The optimal number of clusters would be 4, as it is selected by 20 indices among 30, which is the correct number of clusters. The second option consists in considering only indices that performed best in simulation studies. For example, the 5 top performers in the Milligan and Cooper (1985) study are CH index, Duda index, Cindex, Gamma and Beale.

4.2. Real data set

In the following, we consider the famous iris (Fisher 1936) data set. This data set consists of samples from each of three species of Iris (Iris “setosa”, Iris “virginica” and Iris “versicolor”). Four features were measured for each sample: the length and the width of the sepals and petals, in centimeters. The data set is composed of 3 classes of 50 instances each, where each class refers to a type of iris plant. One class contains “Iris setosa” and is linearly separable from the other 2; the latter contain both “Iris virginica” and “Iris versicolor” and are not linearly separable from each other. Figure 4 shows the pairwise scatter for the original Iris data.

NbClust includes an option to use a user defined dissimilarity matrix. Clustering results of Iris data set with the argument `diss` set to `"diss_matrix"`, the name of the dissimilarity matrix, the argument `index` set to `"alllong"` and the argument `method` set to `"complete"` are presented below.

```
R> data <- iris[, -5]
R> diss_matrix <- dist(data, method = "euclidean", diag = FALSE)
R> NbClust(data, diss = diss_matrix, distance = "NULL", min.nc = 2,
+    max.nc = 10, method = "complete", index = "alllong")
```

According to the majority rule, 3 would be the best number of clusters in the Iris data set. In fact, 14 among 30 indices select 3 as the optimal number of clusters. However, if we look at the top 5 indices in the Milligan and Cooper (1985) study, only the Cindex and the Beale index select the correct number of clusters in Iris data set. Hence, the majority rule seems to be a more reliable solution for selecting the best number of clusters mainly in the case of real data sets.

```
*** : The Hubert index is a graphical method of determining the number of
      clusters.
      In the plot of Hubert index, we seek a significant knee that
      corresponds to a significant increase of the value of the measure i.e
      the significant peak in Hubert index second differences plot.

*** : The D index is a graphical method of determining the number of
      clusters.
      In the plot of D index, we seek a significant knee (the significant
      peak in Dindex second differences plot) that corresponds to a
      significant increase of the value of the measure.

All 150 observations were used.

*******************************************************************
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 15 proposed 3 as the best number of clusters
* 6 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 3 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
$All.index
        KL       CH Hartigan     CCC     Scott  Marriot    TrCovW   TraceW
2   1.9652 280.8392 240.7478 30.4441  933.9084 977604.0 6868.5401 235.1531
3   5.3598 485.9050  68.8363 35.8668 1210.7629 347351.8  304.1791  89.5250
4  54.0377 495.1816  16.4167 35.6036 1346.7582 249402.3  135.7432  60.9730
5   0.0263 414.3925  51.1371 33.0698 1387.9419 296129.2  121.5044  54.8099
6   7.1653 455.4931  16.8076 33.9870 1506.5585 193380.9   96.9908  40.5198
```

```
7   0.5308 423.7198  20.2960 32.9063 1560.0089 184311.4   93.2005  36.2847
8   2.4071 414.7146   4.4653 32.4873 1628.7974 152185.5   60.9393  31.7749
9   6.5604 372.2046   8.2537 31.0319 1646.9164 170694.1   55.3030  30.8062
10  0.2708 348.6421   9.1553 30.1191 1680.9385 167969.1   55.2821  29.1026
    Friedman    Rubin Cindex     DB Silhouette   Duda Pseudot2   Beale
2   715.2826  40.5663 0.3723 0.7027     0.5160 0.1460 444.4821 13.9360
3   804.1705 106.5545 0.3163 0.7025     0.5136 0.5582  55.4060  1.8840
4   955.5312 156.4512 0.3465 0.7289     0.4998 0.5932  32.9134  1.6216
5   991.9852 174.0431 0.3758 0.9838     0.3462 0.5452  48.3914  1.9801
6  1070.1736 235.4228 0.4032 1.0524     0.3382 0.5656  19.9691  1.7855
7  1171.9307 262.9011 0.3982 1.0030     0.3298 0.6480  19.5552  1.2760
8  1251.1704 300.2146 0.4118 1.0738     0.3240 2.1863 -11.9371 -1.2530
9  1290.8832 309.6552 0.4098 0.9954     0.3258 0.6340   5.7720  1.2668
10 1353.2708 327.7814 0.4045 1.0396     0.3095 0.6575   9.8984  1.1948
   Ratkowsky     Ball Ptbiserial     Gap     Frey McClain  Gamma    Gplus
2     0.4729 117.5765     0.6369 -0.2356   0.2675  0.4228 0.7472 353.1090
3     0.4922  29.8417     0.7203  0.1343   0.8589  0.4964 0.8928 139.9284
4     0.4387  15.2432     0.6948 -0.1465 134.6913  0.5734 0.9261  87.9342
5     0.4026  10.9620     0.6073 -0.3669   1.1448  0.7936 0.8589 149.0951
6     0.3738   6.7533     0.5295 -0.3256   0.6883  1.0742 0.8919  88.5252
7     0.3482   5.1835     0.5212 -0.5714   1.2624  1.1037 0.9020  77.1718
8     0.3275   3.9719     0.4753 -0.6911   0.5934  1.3191 0.9115  58.7781
9     0.3092   3.4229     0.4729 -0.9371   0.7370  1.3284 0.9145  56.0378
10    0.2941   2.9103     0.4688 -1.1656   0.7430  1.3469 0.9179  52.7862
        Tau   Dunn Hubert SDindex Dindex   SDbw
2  2475.495 0.0824 0.0015  1.8326 1.1446 0.8976
3  2649.840 0.1033 0.0020  1.6226 0.6722 0.2350
4  2495.851 0.1365 0.0022  1.9103 0.5832 0.1503
5  2206.153 0.1000 0.0022  3.4597 0.5513 0.5055
6  1728.103 0.1311 0.0023  3.5342 0.4778 0.3126
7  1664.993 0.1346 0.0023  3.6106 0.4530 0.2284
8  1384.061 0.1529 0.0023  3.9101 0.4239 0.0357
9  1367.483 0.1539 0.0023  4.0152 0.4171 0.0312
10 1340.581 0.1543 0.0024  4.0261 0.4060 0.0303

$All.CriticalValues
   CritValue_Duda CritValue_PseudoT2 Fvalue_Beale CritValue_Gap
2          0.6121            48.1694       0.0000       -0.3642
3          0.6027            46.1391       0.1134        0.2891
4          0.5551            38.4707       0.1704        0.2300
5          0.5800            42.0003       0.0983       -0.0305
6          0.4590            30.6444       0.1373        0.2590
7          0.5131            34.1652       0.2822        0.1346
8          0.4284            29.3527       1.0000        0.2635
9          0.2576            28.8239       0.2990        0.2483
10         0.3999            28.5079       0.3200        0.2200
```

```
$Best.nc
                     KL       CH Hartigan      CCC    Scott  Marriot
Number_clusters  4.0000   4.0000   3.0000  3.0000   3.0000       3.0
Value_Index     54.0377 495.1816 171.9115 35.8668 276.8545 532302.7
                  TrCovW  TraceW Friedman    Rubin Cindex     DB
Number_clusters    3.000   3.000   4.0000   6.0000 3.0000 3.0000
Value_Index     6564.361 117.076 151.3607 -33.9014 0.3163 0.7025
                Silhouette   Duda PseudoT2 Beale Ratkowsky    Ball
Number_clusters      2.000 4.0000   4.0000 3.000    3.0000  3.0000
Value_Index          0.516 0.5932  32.9134 1.884    0.4922 87.7349
                PtBiserial    Gap Frey McClain  Gamma   Gplus     Tau
Number_clusters     3.0000 3.0000    1  2.0000 4.0000 10.0000    3.00
Value_Index         0.7203 0.1343   NA  0.4228 0.9261 52.7862 2649.84
                   Dunn Hubert SDindex Dindex    SDbw
Number_clusters 10.0000      0  3.0000      0 10.0000
Value_Index      0.1543      0  1.6226      0  0.0303

$Best.partition
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 2 3 2 3 2 3 3 3 3 2 3 2 3 3 2 3
 [71] 2 3 2 2 2 2 2 2 2 3 3 3 3 2 3 2 2 2 3 3 3 2 3 3 3 3 3 2 3 3 2 2 2 2 2
[106] 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[141] 2 2 2 2 2 2 2 2 2 2
```

5. Conclusion

In this paper, we describe clustering methods and validity indices implemented in the **NbClust** package. One major advantage of this new package is that it provides an exhaustive list of indices which to a large extent have not been implemented before in any R package. The current version of **NbClust** contains up to 30 indices. It enables the user to simultaneously vary the number of clusters, the clustering method and the indices to decide how best to group the observations in a data set. Moreover, for each index, **NbClust** proposes the best number of clusters from the different results. The user can thus compare all indices and clustering methods.

Lastly, implementing the validation measures within the R package **NbClust** provides the additional advantage that it can interface with numerous other clustering algorithms in existing R packages. Hence, the **NbClust** package is a valuable addition to the growing collection of cluster validation software available for researchers.

As with many other software packages, **NbClust** package is continually being augmented and improved. We are currently investigating other possible solutions for the final choice of the best number of clusters, such as the building of a composite index and the consideration of sensibility analysis, see Saisana, Saltelli, and Tarantola (2005) and Marozzi (2014), by combining well-chosen indices into a single index. Another future direction includes expanding the functionality of **NbClust** to allow for applying other clustering algorithms such as self organizing maps.

Acknowledgments

The authors would like to thank the NSERC-Industrial Alliance Chair for Women in Science and Engineering in Quebec for the support to this research.

References

Baker FB, Hubert LJ (1975). "Measuring the Power of Hierarchical Cluster Analysis." *Journal of the American Statistical Association*, **70**(349), 31–38.

Ball GH, Hall DJ (1965). "ISODATA: A Novel Method of Data Analysis and Pattern Classification." Stanford Research Institute, Menlo Park. (NTIS No. AD 699616).

Beale EML (1969). *Cluster Analysis*. Scientific Control Systems, London.

Bezdek JC, Pal NR (1998). "Some New Indexes of Cluster Validity." *IEEE Transactions on Systems, Man and Cybernetics*, **28**(3), 301–315.

Brock G, Pihur V, Datta S (2014). ***clValid**: Validation of Clustering Results*. R package version 0.6-6, URL http://CRAN.R-project.org/package=clValid.

Brock G, Pihur V, Datta S, Datta S (2008). "**clValid**: An R Package for Cluster Validation." *Journal of Statistical Software*, **25**(4), 1–22. URL http://www.jstatsoft.org/v25/i04/.

Calinski T, Harabasz J (1974). "A Dendrite Method for Cluster Analysis." *Communications in Statistics – Theory and Methods*, **3**(1), 1–27.

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014). ***NbClust** Package for Determining the Best Number of Clusters*. R package version 2.0.3, URL http://CRAN.R-project.org/package=NbClust.

Chipman H, Tibshirani R (2014). ***hybridHclust**: Hybrid Hierarchical Clustering*. R package version 1.0.4, URL http://CRAN.R-project.org/package=hybridHclust.

Davies DL, Bouldin DW (1979). "A Cluster Separation Measure." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**(2), 224–227.

Dimitriadou E (2014). ***cclust**: Convex Clustering Methods and Clustering Indexes*. R package version 0.6-18, URL http://CRAN.R-project.org/package=cclust.

Dimitriadou E, Dolnicar S, Weingessel A (2002). "An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets." *Psychometrika*, **67**(3), 137–160.

Duda RO, Hart PE (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

Dunn J (1974). "Well Separated Clusters and Optimal Fuzzy Partitions." *Journal Cybernetics*, **4**(1), 95–104.

Edwards AWF, Cavalli-Sforza L (1965). "A Method for Cluster Analysis." *Biometrics*, **21**(2), 362–375.

Everitt B (1974). *Cluster Analysis.* Heinemann Educational, London.

Fisher RA (1936). "The Use of Multiple Measurements in Taxonomic Problems." *The Annals of Eugenics*, **7**(2), 179–188.

Florek K, Lukaszewicz J, Perkal J, Zubrzycki S (1951). "Sur la Liaison et la Division des Points d'un Ensemble Fini." *Colloquium Mathematicae*, **2**(3–4), 282–285.

Frey T, Van Groenewoud H (1972). "A Cluster Analysis of the D-Squared Matrix of White Spruce Stands in Saskatchewan Based on the Maximum-Minimum Principle." *Journal of Ecology*, **60**(3), 873–886.

Friedman HP, Rubin J (1967). "On Some Invariant Criteria for Grouping Data." *Journal of the American Statistical Association*, **62**(320), 1159–1178.

Fukunaga K, Koontz WLG (1970). "A Criterion and An Algorithm for Grouping Data." *IEEE Transactions on Computers*, **C-19**(10), 917–923.

Gordon AD (1999). *Classification.* 2nd edition. Chapman & Hall/CRC, London.

Gower JC (1967). "A Comparison of Some Methods of Cluster Analysis." *Biometrics*, **23**(4), 623–637.

Halkidi M, Batistakis I, Vazirgiannis M (2001). "On Clustering Validation Techniques." *Journal of Intelligent Information Systems*, **17**(2/3), 107–145.

Halkidi M, Vazirgiannis M (2001). "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set." In *ICDM'01 Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 187–194.

Halkidi M, Vazirgiannis M, Batistakis I (2000). "Quality Scheme Assessment in the Clustering Process." In *Principles of Data Mining and Knowledge Discovery*, volume 1910 of *Lecture Notes in Computer Science*, pp. 265–276. Springer-Verlag, Berlin Heidelberg. Proceedings of the 4th European Conference, PKDD 2000, Lyon, France, September 13–16 2000.

Hartigan JA (1975). *Clustering Algorithms.* John Wiley & Sons, New York.

Hartigan JA, Wong MA (1979). "A K-Means Clustering Algorithm." *Journal of the Royal Statistical Society C*, **28**(1), 100–108.

Hill RS (1980). "A Stopping Rule for Partitioning Dendrograms." *Botanical Gazette*, **141**(3), 321–324.

Hornik K (2005). "A CLUE for CLUster Ensembles." *Journal of Statistical Software*, **14**(12), 1–25. URL http://www.jstatsoft.org/v14/i12/.

Hornik K (2014). ***clue**: Cluster Ensembles.* R package version 0.3-48, URL http://CRAN.R-project.org/package=clue.

Hubert LJ, Arabie P (1985). "Comparing Partitions." *Journal of Classification*, **2**(1), 193–218.

Hubert LJ, Levin JR (1976). "A General Statistical Framework for Assessing Categorical Clustering in Free Recall." *Psychological Bulletin*, **83**(6), 1072–1080.

Jain AK, Murty PJ, Flyn PJ (1998). "Data Clustering: A Review." *ACM Computing Surveys*, **31**(3), 264–323.

Kaufman L, Rousseeuw PJ (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Kraemer HC (1982). *Biserial Correlation*. John Wiley & Sons. Reference taken from a SAS note about the BISERIAL macro on this Web Site: http://support.sas.com/kb/24/991.html.

Krzanowski WJ, Lai YT (1988). "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering." *Biometrics*, **44**(1), 23–34.

Lebart L, Morineau A, Piron M (2000). *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

MacQueen JB (1967). "Some Methods for Classification and Analysis of Multivariate Observations." In LML Cam, J Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297.

Maechler M, Rousseeuw P, Struyf A, Hubert M (2014). ***cluster**: Cluster Analysis Extended Rousseeuw et al.* R package version 1.15.2, URL http://CRAN.R-project.org/package=cluster.

Marozzi M (2014). "Construction, Dimension Reduction and Uncertainty Analysis of an Index of Trust in Public Institutions." *Quality and Quantity*, **48**(2), 939–953.

Marriot FHC (1971). "Practical Problems in a Method of Cluster Analysis." *Biometrics*, **27**(3), 501–514.

McClain JO, Rao VR (1975). "**CLUSTISZ**: A Program to Test for The Quality of Clustering of a Set of Objects." *Journal of Marketing Research*, **12**(4), 456–460.

McQuitty LL (1966). "Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data." *Educational and Psychological Measurement*, **26**(4), 825–831.

Milligan GW (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika*, **45**(3), 325–342.

Milligan GW (1981). "A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis." *Psychometrika*, **46**(2), 187–199.

Milligan GW, Cooper MC (1985). "An Examination of Procedures for Determining the Number of Clusters in a Data Set." *Psychometrika*, **50**(2), 159–179.

Murtagh F, Legendre P (2011). "Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm." Unpublished preprint.

Murtagh F, Legendre P (2014). "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" *Journal of Classification*. Forthcoming.

Nieweglowski L (2014). ***clv**: Cluster Validation Techniques*. R package version 0.3-2.1, URL http://CRAN.R-project.org/package=clv.

Orloci L (1967). "An Agglomerative Method for Classification of Plant Communities." *Journal of Ecology*, **55**(1), 193–206.

Ratkowsky DA, Lance GN (1978). "A Criterion for Determining the Number of Groups in a Classification." *Australian Computer Journal*, **10**(3), 115–117.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rohlf FJ (1974). "Methods of Comparing Classifications." *Annual Review of Ecology and Systematics*, **5**, 101–113.

Rousseeuw P (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Saisana M, Saltelli A, Tarantola S (2005). "Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators." *Journal of the Royal Statistical Society A*, **168**(2), 307–323.

Sarle WS (1983). "SAS Technical Report A-108, Cubic Clustering Criterion." SAS Institute Inc. Cary, NC.

SAS Institute Inc (2012). *SAS/STAT Software, Version 12.1*. SAS Institute Inc., Cary, NC. URL http://www.sas.com/.

Scott AJ, Symons MJ (1971). "Clustering Methods Based on Likelihood Ratio Criteria." *Biometrics*, **27**(2), 387–397.

Seber GAF (1984). *Multivariate Observations*. John Wiley & Sons, New York.

Sheikholeslami C, Chatterjee S, Zhang A (2000). "WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Database." *The International Journal on Very Large Data Bases*, **8**(3–4), 289–304.

Sokal R, Michener C (1958). "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Science Bulletin*, **38**(22), 1409–1438.

Sørensen TA (1948). "A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and its Application to Analyses of the Vegetation on Danish Commons." *Biologiske Skrifter*, **5**, 1–34.

Suzuki R, Shimodaira H (2014). ***pvclust**: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. R package version 1.2-2, URL http://CRAN.R-project.org/package=pvclust.

Templ M (2007). ***clustTool**: GUI for Clustering Data with Spatial Information*. R package version 1.3, URL http://CRAN.R-project.org/package=clustTool.

Theodoridis S, Koutroubas K (2008). *Pattern Recognition*. 4th edition. Academic Press.

Tibshirani R, Walther G, Hastie T (2001). "Estimating the Number of Clusters in a Data Set Via the Gap Statistic." *Journal of the Royal Statistical Society B*, **63**(2), 411–423.

Walesiak M, Dudek A (2014). ***clusterSim**: Searching for Optimal Clustering Procedure for a Data Set.* R package version 0.43-4, URL http://CRAN.R-project.org/package=clusterSim.

Ward JH (1963). "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association*, **58**(301), 236–244.

Affiliation:

Malika Charrad
Université de Gabes
Institut Supérieur de l'Informatique
Route Djerba Km 3, Boite Postale N 283
4100 Medenine, Tunisie
and
Université Laval, Québec
E-mail: malika.charrad@riadi.rnu.tn

Nadia Ghazzali
Université du Québec à Trois-Rivières
E-mail: nadia.ghazzali@uqtr.ca

Véronique Boiteau, Azam Niknafs
Département de Mathématiques et de Statistique
Université Laval, Québec
E-mail: veronique.boiteau.1@ulaval.ca, azam.niknafs.1@ulaval.ca

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 61, Issue 6 *Submitted:* 2012-08-13
October 2014 *Accepted:* 2013-04-08