Determining the Number of Groups in Cluster Analysis Using Classical Indexes and Stability Measures—Comparison of Results

Dorota Rozmus

Abstract In the context of taxonomic methods, in recent years, much attention has been paid to the issue of the stability of these methods, i.e., the answer to the question: to what extent the structure discovered by a given method is actually present in the data. This criterion examines whether the groups that were created as a result of using clustering method to a set of objects are real (the structure is stable), or whether they appeared accidentally. Most often this criterion is used when selecting the number of groups (k), for which should be clustered a set of data. The aim of the article is to compare the results in terms of the indicated correct number of groups by classical indexes and stability measures.

Keywords Clustering · Stability measures · Internal indexes

1 Introduction

The main problem in taxonomy is to determine whether the groups that we received reflect the actual structure of general population (which generated the data). This involves the problem of "clustering model" identification, e.g., the number of groups k, a distance metric, the control parameters of an algorithm. Recently, the stability criterion increasingly gains in popularity in response to these problems.

Informally, this criterion states that if a cluster algorithm is repeatedly used for independent samples of objects (with unchanged parameters of the algorithm), resulting in similar grouping results, it can be considered as stable and reflecting the actual structure of the groups (Shamir and Tishby 2008). Volkovich et al. (2010) even state that the number of groups that maximizes the stability of clustering can serve as an estimate of the "true" number of groups.

The literature proposes a number of different ways for measuring stability (e.g., Ben-Hur and Guyon 2003; Brock et al. 2008; Henning 2007; Fang and Wang 2012;

D. Rozmus (✉)
University of Economics in Katowice, Katowice, Poland
e-mail: dorota.rozmus@ue.katowice.pl

K. Jajuga et al. (eds.), *Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-75190-6_2

Lord et al. 2017; Marino and Presti 2019; Suzuki and Shimodaira 2006). Theoretical considerations have also led to the development of computer tools for the practical implementation of the proposed ways to study stability. The practical tools are available within several **R** packages, for example: `clv`, `clValid`, `ClusterStability`, `fpc`, `pvclust`.

Due to the hypothesis that the stability of clusters may be the answer to the question about the appropriate number of clusters (k), the aim of the article will be to compare the results in the context of the indicated value of k by classical indexes that so far served this issue (e.g., Hubert and Levin index, Dunn index, Silhouette index) and the cluster stability measures proposed in the literature.

2 Measures of Cluster Stability

This part of the article presents the research methods, i.e., cluster stability measures. In this study, only such stability measures that one can find in the **R** program were used, i.e., measures from packages: `clv`, `clValid` and `fpc`.[1] There are much more packages for stability testing of course, but those mentioned libraries can be used with various clustering methods, e.g., k-means, k-medoids, hierarchical, and others.

As the classical indexes for determining the number of groups in clutering are well known, they will not be discussed in details. It should be mentioned, however, that only internal measures will be used.

2.1 Ben-Hur and Guyon Stability Measure

The concept of stability by Ben-Hur and Guyon (2003) is based on the finding that if the clustering properly represents the structure in the data, it should be stable with respect to small changes in the data set. They proposed two measures of stability: a measure based on the index of similarity between two partitions[2] and a measure based on the pattern-wise agreement concept[3].

The algorithm of calculating of stability measure based on the index of similarity between two partitions can be described in the following steps:

1. Cluster the original data set in order to obtain the reference partition.
2. Select a random subsample of observations from the original data set and group the objects from this subsample.

[1]Packages `clv`, `clValid` and `fpc`, were also selected because the methods implemented there have been the subject of the author's research for a long time (e.g., Rozmus 2017).

[2]This measure is implemented by the function `cls.stab.sim.ind` in `clv` package in **R**.

[3]This measure is implemented by the function `cls.stab.opt.assign` in `clv` package in **R**.

3. Calculate the stability between the reference partition and the partition of the subsample using the index of similarity between two partitions (e.g., Rand index).
4. Repeat the procedure several times.
5. Repeat the procedure for different values of k (number of groups).

The pattern-wise agreement concept of stability measure is based on the idea of pattern-wise agreement and pattern-wise stability.

Given two groupings L_1 and L_2, pattern-wise agreement can be defined as follows:

$$\delta_\sigma(i) = \begin{cases} 1, & if\ \sigma(L_1(i)) = L_2(i), \\ 0, & if\ \sigma(L_1(i)) \neq L_2(i), \end{cases} \tag{1}$$

where $\sigma : \{1, \ldots, k_1\} \rightarrow \{1, \ldots, k_2\}$.

Pattern-wise stability is defined as the fraction of subsampled partitions where the subsampled labeling of observation i agrees with that of the reference labeling, by averaging the pattern-wise agreement:

$$\mathrm{n(i)} = \frac{1}{\mathrm{N_i}} \sum \delta_\sigma(\mathrm{i}) \tag{2}$$

where N_i—number of subsamples where pattern i appears.

The stability of group j in the reference partition is the average of pattern-wise stability:

$$c(j) = \frac{1}{|L_1 = j|} \sum\nolimits_{i \in (L_1 = j)} n(i) \tag{3}$$

where $|-|$ means cardinality of the set.

The stability of the reference partition into k groups is defined as:

$$S_k = \min_j c(j). \tag{4}$$

Finally, the most stable clustering is indicated by the maximum of S_k.

2.2 Brock, Pihur, Datta, and Datta Stability Measure

Measures of stability by Brock et al. (2008)[4] are dedicated mainly for validating the results of clustering analysis in biology. There are three main types of cluster validation measures available: "internal," "biological," and "stability."

[4]This measure can be found in `clValid` package in **R**.

The article focuses only on the last group of measures. They evaluate the stability of a clustering result by comparing it with the clusters obtained by removing one column (i.e., variable) at a time (Brock et al. 2008). These measures include: the average proportion of non-overlap (APN), the average distance (AD), the average distance between means (ADM), and the figure of merit (FOM).

Only APN was used in experiments because this is the only measure that is normalized in the interval (0, 1), with values close to zero corresponding with highly consistent clustering results. APN measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed:

$$APN = \frac{1}{M \cdot N} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{M} \left(1 - \frac{n(C^{i,j} \cap C^{i,0})}{n(C^{i,0})}\right), \tag{5}$$

where

$C^{i,0}$ represents the cluster containing observation i using the original clustering (based on all available data),
$C^{i,j}$ represents the cluster containing observation i where the clustering is based on the data set with j column removed,
$n(\cdot)$ is the cardinality of a cluster,
N denotes the total number of observations (rows) in a data set,
M denotes the total number of variables (columns) in a data set.

2.3 Fang and Wang Stability Measure

Fang and Wang stability measures (2012)[5] focus on the concept of stability as robustness to randomness present in the sample. Drawing on the work of Wang (2010), they formulate the concept of stability in the following way: if one draws samples from the population and applies a selected clustering algorithm, the results of grouping should not be very different.

Presented Fang and Wang measure is based on the following general idea: Several times two bootstrap samples are drawn from the data, and the number of clusters is chosen by optimizing an instability estimation from these pairs.

Denoting a cluster algorithm with $k \geq 2$ groups by $\Psi(\cdot, k)$, when we use it to sample X^n, we get the clustering $\Psi_{X^n,k}(x)$; the algorithm can be presented according to the following procedure. For the assumed value of $k = 2, \ldots, K$:

[5]This measure can be found in fpc package in **R**. It includes two functions for measuring stability: clusterboot and nselectboot. In the experiments only the nselectboot function was used.

1. Construct B independent pairs of bootstrap samples $\left(X_b^{n*}, \tilde{X}_b^{n*}\right)$, $b = 1, \ldots, B$.
2. Make groupings $\Psi_{X_b^{n*},k}$ and $\Psi_{\tilde{X}_b^{n*},k}$ on $\left(X_b^{n*}, \tilde{X}_b^{n*}\right)$, $b = 1,\ldots, B$.
3. For each pair, $\Psi_{X_b^{n*},k}$ and $\Psi_{\tilde{X}_b^{n*},k}$ calculate the empirical clustering distance:

$$d\left(\Psi_{X_b^{n*},k}, \Psi_{\tilde{X}_b^{n*},k}\right) = \\ = \frac{1}{n^2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}\left|I\left\{\Psi_{X_b^{n*},k}(x_i) = \Psi_{X_b^{n*},k}(x_j)\right\} - I\left\{\Psi_{\tilde{X}_b^{n*},k}(x_i) = \Psi_{\tilde{X}_b^{n*},k}(x_j)\right\}\right|. \tag{6}$$

4. Instability of clustering is calculated as:

$$\hat{s}_B = \frac{1}{B}\sum\nolimits_{b=1}^{B} d\left(\Psi_{X_b^{n*},k}, \Psi_{\tilde{X}_b^{n*},k}\right). \tag{7}$$

3 A Data Set and the Scheme of Research

A data set was built on the data obtained from the sustainable development indicators application developed by the Central Statistical Office in Poland. This application monitors the implementation of the sustainable development policy in the EU countries. The data are divided into four groups, monitoring the implementation of the sustainable development policy within the following domains:

- social,
- economic,
- environmental,
- institutional and political.

The study used the data from 2015, comprising 19 metric variables in the social domain, 18 variables in the economic domain, 11 in the environmental domain, and 15 in the institutional and political domain (only complete data were used).

Clustering was carried out within each domain separately. This is related to the idea of weak and strong sustainability (Borys 2005, 2014; Lorek 2011). In accordance with weak sustainability, it is permissible to consider all domains together, because the resources from these domains are considered substitutable. According to strong sustainability, resources within each domain are considered to be complementary, and therefore, every order should be considered separately because it is not possible to develop one domain at the expense of the other. And in this spirit, the analysis presented in the paper was carried out.

As a clustering methods, two partitioning algorithms were used, i.e., *k*-means and *k*-medoids and two hierarchical algorithms, i.e., group average linkage and Ward method.

Among the classical indexes used to determine the number of clusters, only internal measures was chosen, i.e., Hubert and Levin index, Davies and Bouldin index, CalińskiHarabasz index, and Silhouette index. As these are commonly known and recognized measures, they will not be discussed in detail in the paper[6].

In Ben-Hur and Guyon measure of stability, similarity between two partitions were tested with all available in `clv` package indices, i.e., Rand, dot product, similarity index, and Jaccard. For creating subsamples, the subset ratio was equal 0.8.

In stability measure proposed by Fang and Wang (`fpc` package), 100 bootstrap samples were created. Each new data set (of the same size as the original) is created by resampling the original data set with replacement.

4 Empirical Results

As it was mentioned before, the study was carried within each domain separately. The results are discussed below.

In Tables 2, 3, 4, 5, there are presented values of classical indexes and different stability measures used in the experiments. The measures were calculated only for $k = 2, \ldots, 5$. Abbreviations used on those figures are explained in Table 1. In last column, there is presented the information about optimization direction of the criterion.

4.1 Results for the Social Domain

Looking at the values for different stability measures (Table 2), it can be seen that classical indexes and stability measures suggest different value of k (number of groups). Indexes in most cases point $k = 2$ as the optimal value, whereas stability measures the most often indicate $k = 3$. It is also worth paying attention to the Hubert and Levin and Dunn index, which, like Fang and Wang stability measure (FW), suggest the maximum number of groups under consideration. This scheme will also appear in other domains (especially for Dunn index and Fang and Wang stability measure).

[6] The indexes were calculated using the functions from the `clusterSim` and `clusterCrit` packages.

Table 1 Description of abbreviations for the stability measures used in the presentation of the results and the direction of their optimization

Abbreviation	Description	Optimization direction
BH-G rand	Ben-Hur and Guyon measure of stability, with Rand similarity index (implemented by the function `cls.stab.sim.ind`)	max
BH-G dot	Ben-Hur and Guyon measure of stability, with dot product similarity index (implemented by the function `cls.stab.sim.ind`)	max
BH-G sim	Ben-Hur and Guyon measure of stability, with similarity index (implemented by the function `cls.stab.sim.ind`)	max
BH-G jaccard	Ben-Hur and Guyon measure of stability, with Jaccard similarity index (implemented by the function `cls.stab.sim.ind`)	max
BH-G1	Ben-Hur and Guyon measure of stability implemented by the function `cls.stab.sim.opt.assigned`	max
B et al.	Measures of stability by Brock et al. indicated by the average proportion of non-overlap	min
FW	Fang and Wang stability measure (implemented by `nselectboot` function)	min

Source Own computations

4.2 Results for the Economic Domain

Looking at the results for the economic domain, it can be generally stated that the classical indexes and measures of stability are, in most cases, compatible, suggesting $k = 2$. The only exceptions are the average method, where the indexes indicate $k = 5$ as correct, while the stability measures indicate $k = 2$. Moreover, it can be observed that for most of the considered clustering methods, Dunn index and Fang and Wang stability measure (FW) suggest the maximum considered number of groups.

4.3 Results for the Environmental Domain

A huge divergence of results as to the indications of the actual number of groups can be observed for this domain. For example, for the k-means method, the indexes most often suggest $k = 4$, while the stability measures indicate very different values (from $k = 2$ to $-k = 5$). A very large variation in the suggested value of the k parameter can be noticed for the k-medoids method, both in the context of classical indexes and cluster stability measures. For hierarchical methods, the indices and

Table 2 Results of clustering for partitioning method (social domain)

Index					Stability measures				
k-means	**#2**	**#3**	**#4**	**#5**	**k-means**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.167	0.151	0.112	**0.102**	BH-G rand	0.856	**0.954**	0.907	0.843
Davies and Bouldin	**0.848**	1.526	1.526	1.526	BH-G dot	0.862	**0.928**	0.816	0.651
Caliński-Harabasz	**36.170**	20.356	9.820	9.658	BH-G sim	0.845	**0.940**	0.806	0.629
Silhouette	**0.247**	0.200	0.186	0.183	BH-G jaccard	0.659	**0.897**	0.673	0.519
Dunn	0.388	0.437	0.456	**0.509**	BH-G1	0.855	**0.962**	0.806	0.429
					B et al.	**0.021**	0.025	0.045	0.055
					FW	0.110	0.065	0.067	**0.062**
k-medoids	**#2**	**#3**	**#4**	**#5**	**k-medoids**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.190	0.195	0.178	**0.158**	BH-G rand	0.749	**0.975**	0.914	0.958
Davies and Bouldin	**0.692**	1.787	4.116	4.261	BH-G dot	0.760	**0.959**	0.816	0.884
Caliński-Harabasz	23.754	**31.808**	23.437	17.629	BH-G sim	0.727	**0.974**	0.791	0.894
Silhouette	**0.229**	0.149	0.133	0.109	BH-G jaccard	0.814	**0.883**	0.588	0.84
Dunn	**0.415**	0.318	0.328	0.328	BH-G1	0.918	**0.958**	0.913	0.765
					B et al.	0.037	0.034	**0.033**	0.077
					FW	0.147	0.145	0.135	**0.117**
Average	**#2**	**#3**	**#4**	**#5**	**Average**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.237	0.163	0.146	**0.145**	BH-G rand	0.809	0.789	**0.899**	0.891
Davies and Bouldin	**0.827**	1.142	1.987	1.270	BH-G dot	**0.836**	0.747	0.828	0.765
Caliński-Harabasz	1.358	14.523	9.927	8.654	BH-G sim	0.805	0.673	**0.875**	0.742
Silhouette	**0.236**	0.177	0.150	0.129	BH-G jaccard	0.595	0.682	**0.706**	0.674
Dunn	**0.445**	0.408	0.408	0.408	BH-G1	0.387	0.651	**0.884**	0.437

(continued)

Table 2 (continued)

Index					Stability measures				
					B et al.	0.052	**0.025**	0.027	2,000
					FW	0.212	0.214	0.199	**0.174**
Ward	**#2**	**#3**	**#4**	**#5**	**Ward**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.199	0.174	0.149	**0.100**	BH-G rand	0.772	0.902	**0.914**	0.909
Davies and Bouldin	**0.933**	3.362	4.639	3.814	BH-G dot	0.787	**0.845**	0.821	0.768
Caliński-Harabasz	**28.842**	14.638	9.638	8.150	BH-G sim	0.685	**0.92**	0.881	0.868
Silhouette	**0.229**	0.144	0.165	0.178	BH-G jaccard	0.583	**0.783**	0.600	0.582
Dunn	0.366	0.429	0.429	**0.509**	BH-G1	0.784	**0.847**	0.763	0.708
					B et al.	0.044	**0.034**	0.041	0.047
					FW	0.128	0.130	0.131	**0.119**

# denotes number of clusters
Source Own computations

Table 3 Results of clustering for partitioning method (economic domain)

Index					Stability measures				
k-means	**#2**	**#3**	**#4**	**#5**	**k-means**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.201	0.184	**0.123**	0.128	BH-G rand	0.871	0.883	**0.920**	0.894
Davies and Bouldin	**1.079**	1.248	1.084	1.413	BH-G dot	**0.886**	0.845	0.850	0.743
Caliński-Harabasz	**22.207**	18.320	20.346	15.946	BH-G sim	**0.886**	0.833	0.820	0.724
Silhouette	**0.227**	0.192	0.207	0.173	BH-G jaccard	**0.811**	0.753	0.765	0.618
Dunn	0.190	0.336	0.382	**0.409**	BH-G1	**0.847**	0.817	0.678	0.472
					B et al.	0.018	**0.017**	0.037	0.039
					FW	**0.051**	0.106	0.092	0.079
k-medoids	**#2**	**#3**	**#4**	**#5**	**k-medoids**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.201	0.176	0.133	**0.124**	BH-G rand	**0.942**	0.879	0.865	0.898
Davies and Bouldin	**0.857**	1.628	1.464	1.292	BH-G dot	**0.949**	0.832	0.776	0.816
Caliński-Harabasz	**21.065**	13.811	9.638	10.969	BH-G sim	**0.949**	0.858	0.739	0.851
Silhouette	**0.227**	0.200	0.202	0.184	BH-G jaccard	**0.9136**	0.742	0.641	0.702
Dunn	0.190	0.278	0.362	**0.372**	BH-G1	**0.986**	0.838	0.542	0.821
					B et al.	**0.029**	0.033	0.035	0.042
					FW	0.144	0.145	0.118	**0.102**
Average	**#2**	**#3**	**#4**	**#5**	**Average**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.210	0.165	0.165	**0.095**	BH-G rand	**0.942**	0.878	0.896	0.880
Davies and Bouldin	1.305	1.509	1.251	**0.961**	BH-G dot	**0.953**	0.894	0.886	0.852
Caliński-Harabasz	2.426	1.915	1.417	**8.281**	BH-G sim	**0.947**	0.873	0.837	0.838
Silhouette	**0.232**	0.174	0.141	0.204	BH-G jaccard	**0.922**	0.812	0.806	0.752
Dunn	**0.489**	0.469	0.469	0.481	BH-G1	**0.989**	0.982	0.757	0.764

(continued)

Table 3 (continued)

Index					Stability measures				
					B et al.	**0.004**	0.011	0.024	0.039
					FW	0.203	0.181	0.137	**0.105**
Ward	**#2**	**#3**	**#4**	**#5**	**Ward**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.221	0.189	0.145	**0.107**	BH-G rand	0.914	0.877	0.890	**0.944**
Davies and Bouldin	**0.961**	1.276	1.077	1.231	BH-G dot	**0.930**	0.835	0.802	0.877
Caliński-Harabasz	**24.834**	18.338	24.184	19.687	BH-G sim	**0.925**	0.82	0.737	0.838
Silhouette	**0.211**	0.174	0.193	0.185	BH-G jaccard	**0.882**	0.751	0.676	0.787
Dunn	0.313	0.321	0.321	**0.385**	BH-G1	**0.993**	0.846	0.723	0.815
					B et al.	**0.022**	0.035	0.048	0.058
					FW	0.135	0.126	0.119	**0.101**

Source Own computations

Table 4 Results of clustering for partitioning method (environmental domain)

Index					Stability measures				
k-means	**#2**	**#3**	**#4**	**#5**	**k-means**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.292	0.205	0.117	**0.105**	BH-G rand	0.786	0.849	0.806	**0.891**
Davies and Bouldin	1.586	1.529	**1.236**	1.424	BH-G dot	**0.808**	0.801	0.671	0.788
Caliński-Harabasz	10.336	6.631	**14.146**	11.594	BH-G sim	0.767	0.773	0.666	0.787
Silhouette	0.185	0.228	**0.259**	0.226	BH-G jaccard	**0.712**	0.677	0.529	0.661
Dunn	0.274	0.317	**0.400**	**0.400**	BH-G1	0.711	**0.886**	0.333	0.632
					B et al.	0.155	0.060	**0.054**	0.076
					FW	0.157	0.103	0.102	**0.095**
k-medoids	**#2**	**#3**	**#4**	**#5**	**k-medoids**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.332	0.193	0.194	**0.124**	BH-G rand	0.715	0.798	**0.833**	0.821
Davies and Bouldin	**1.527**	1.785	2.016	1.603	BH-G dot	**0.758**	0.748	0.701	0.597
Caliński-Harabasz	**8.565**	4.553	4.893	6.406	BH-G sim	0.662	**0.714**	0.709	0.642
Silhouette	0.181	**0.238**	0.191	0.197	BH-G jaccard	**0.637**	0.630	0.571	0.443
Dunn	0.291	0.392	0.392	**0.494**	BH-G1	0.768	**0.903**	0.846	0.738
					B et al.	**0.055**	0.056	0.119	0.120
					FW	0.179	0.155	0.137	**0.126**
Average	**#2**	**#3**	**#4**	**#5**	**Average**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.181	0.153	0.130	**0.086**	BH-G rand	0.865	**0.937**	0.886	0.703
Davies and Bouldin	1.621	1.355	**1.018**	1.037	BH-G dot	0.917	**0.959**	0.913	0.727
Caliński-Harabasz	1.091	5.291	5.802	**9.753**	BH-G sim	0.857	**0.927**	0.875	0.652
Silhouette	**0.303**	0.234	0.204	0.242	BH-G jaccard	0.861	**0.928**	0.850	0.558
Dunn	0.425	0.400	0.452	**0.557**	BH-G1	**0.915**	0.847	0.729	0.563
					B et al.	**0.009**	0.035	0.063	0.119
					FW	0.189	0.216	0.184	**0.148**
Ward	**#2**	**#3**	**#4**	**#5**	**Ward**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.257	0.193	0.114	**0.086**	BH-G rand	0.823	**0.957**	0.869	0.893
Davies and Bouldin	1.698	1.439	1.262	**1.037**	BH-G dot	0.854	**0.943**	0.770	0.781
Caliński-Harabasz	7.933	5.587	9.683	**9.753**	BH-G sim	0.782	**0.947**	0.725	0.796
Silhouette	0.206	0.238	**0.255**	0.242	BH-G jaccard	0.775	**0.909**	0.642	0.656
Dunn	0.362	0.392	0.494	**0.557**	BH-G1	0.900	**0.948**	0.733	0.760
					B et al.	**0.039**	0.052	0.116	0.153
					FW	0.152	0.147	0.130	**0.117**

Source Own computations

Table 5 Results of clustering for partitioning method (institutional and political domain)

Index					Stability measures				
k-means	**#2**	**#3**	**#4**	**#5**	**k-means**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.117	0.118	**0.077**	0.084	BH-G rand	0.895	0.949	0.831	**1.000**
Davies and Bouldin	**2.043**	2.337	2.659	2.640	BH-G dot	0.897	0.919	0.643	**1.000**
Caliński-Harabasz	**6.223**	4.186	3.490	2.575	BH-G sim	0.878	0.891	0.674	**1.000**
Silhouette	**0.318**	0.221	0.216	0.180	BH-G jaccard	0.819	0.856	0.482	**0.950**
Dunn	0.393	0.438	**0.461**	0.409	BH-G1	**1.000**	0.928	0.702	0.435
					B et al.	**0.030**	0.043	0.041	0.043
					FW	0.162	0.169	0.140	**0.120**
k-medoids	**#2**	**#3**	**#4**	**#5**	**k-medoids**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.127	0.154	0.109	**0.108**	BH-G rand	0.916	0.900	0.863	**0.962**
Davies and Bouldin	**2.166**	3.746	3.901	4.310	BH-G dot	0.902	0.846	0.733	**0.960**
Caliński-Harabasz	**4.813**	2.358	2.811	1.937	BH-G sim	0.904	0.838	0.763	**0.929**
Silhouette	**0.316**	0.170	0.178	0.148	BH-G jaccard	0.833	0.743	0.606	**0.972**
Dunn	0.349	0.355	0.371	**0.413**	BH-G1	**0.989**	0.891	0.904	0.575
					B et al.	**0.029**	0.069	0.054	0.159
					FW	**0.020**	0.052	0.085	0.088
Average	**#2**	**#3**	**#4**	**#5**	**Average**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.137	0.118	0.077	0.069	BH-G rand	**0.960**	0.897	0.926	0.959
Davies and Bouldin	**2.200**	2.747	2.659	2.485	BH-G dot	**0.970**	0.899	0.889	0.955
Caliński-Harabasz	**5.370**	3.547	3.490	2.634	BH-G sim	**0.947**	0.850	0.898	0.923
Silhouette	**0.300**	0.211	0.216	0.192	BH-G jaccard	0.943	0.823	0.807	**0.962**
Dunn	0.458	0.458	**0.461**	**0.461**	BH-G1	0.875	0.844	0.875	**0.925**

(continued)

Table 5 (continued)

Index					Stability measures				
					B et al.	**0.032**	0.034	0.035	0.044
					FW	0.200	0.189	0.170	**0.149**
Ward	**#2**	**#3**	**#4**	**#5**	**Ward**	**#2**	**#3**	**#4**	**#5**
Hubert and Levin	0.107	0.138	0.097 0	**0.782027**	BH-G rand	0.938	0.916	0.873	**0.969**
Davies and Bouldin	2.301	2.082	3.389	**1.639384**	BH-G dot	0.916	0.878	0.782	**0.970**
Caliński-Harabasz	2.934	4.180	3.044	**7.86463**	BH-G sim	0.927	0.862	0.742	**0.942**
Silhouette	**0.356**	0. 217	0.207	0.2108244	BH-G jaccard	0.862	0.801	0.651	**0.876**
Dunn	0. 386	0.375	0.442	**0.461217**	BH-G1	**0.972**	0.657	0.760	0.675
					B et al.	**0.071**	0.078	0.091	0.086
					FW	0.144	0.160	0.143	**0.113**

Source Own computations

measures of stability are more consistent, suggesting that $k = 5$ by index, and $k = 3$ by stability measures.

Nevertheless, despite the large variation in the selected value of the k, one can again observe very similar behavior of Hubert and Levin, Dunn indexes and Fang and Wang stability measure (often suggest the largest number of groups considered).

4.4 Results for the Institutional and Political Domain

In terms of the institutional and political domain, very large discrepancies in the suggested value of the k parameter can be observed for k-means and k-medoids. For k-means, the indexes most often suggest $k = 2$, while the stability measures most often indicate $k = 5$. Similar conclusions can be drawn for k-medoids (although for stability measures, the clustering into two groups is also often correct). For hierarchical methods, the same conclusions as for the k-means can be seen for the average method. On the other hand, for the Ward method, the indexes and stability measures are quite unanimous, showing in most cases the correctness of the clustering into five groups.

Again, it is also worth paying attention to the Dunn index which behaves similarly to Fang and Wang stability measure, suggesting the largest considered number of clusters.

5 Conclusions

Summing up this research, the results of which are presented in this paper, it should be stated that the final result depends on the chosen method. As a rule, very often classical indexes show a different value of the k parameter than the stability measures. Moreover, it can be seen that this value is lower for indexes than for stability measures.

Another summary conclusion resulting from the conducted research is that the Dunn index (and also the Hubert and Levin index in some cases) very often behaves like Fang and Wang stability measure, favoring the clustering into the largest number of groups under consideration. In additional studies (the results of which are not presented here), the value of the parameter k was increased to 10, and this principle was still revealed.

It seems that an interesting topic of future research will be to compare the results using external indexes (e.g., Rand index) and stability measures on benchmark sets with the known cluster structure.

References

Ben-Hur A, Guyon I (2003) Detecting stable clusters using principal component analysis. Methods Mol Biol 224:159–182

Borys T (ed) (2005) Wskaźniki zrównoważonego rozwoju. Wydawnictwo Ekonomia i Środowisko, Warszawa-Białystok

Borys T (2014) Wybrane problemy metodologii pomiaru nowego paradygmatu rozwoju—polskie doświadczenia. Optimum. Studia Ekonomiczne 3(69):3–21

Brock G, Pihur V, Datta S, Datta S (2008) `clValid`: an R package for cluster validation. J Stat Softw 25(4)

Fang Y, Wang J (2012) Selection of the number of clusters via the bootstrap method. Comput Stat Data Anal 56:468–477

Henning C (2007) Cluster-wise assessment of cluster stability. Comput Stat Data Anal 52:258–271

Lord E, Willems M, Lapointe FJ, Makarenkov V (2017) Using the stability of objects to determine the number of clusters in datasets. Inf Sci 393:29–46

Lorek E (2011) Ekonomia zrównoważonego rozwoju w badaniach polskich i niemieckich. In: Kos B (ed) Transformacja gospodarki—poziom krajowy i międzynarodowy. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach 90:103–112

Marino V, Presti LL (2019) Stay in touch! New insights into end-user attitudes towards engagement platforms. J Consum Mark 36:772–783

Rozmus D (2017) Using R packages for comparison of cluster stability. Acta Universitatis Lodziensis Folia Oeconomica 330(4):77–86

Shamir O, Tishby N (2008) Cluster stability for finite samples. Adv Neural Inf Process Syst 20:1297–1304

Suzuki R, Shimodaira H (2006) `Pvclust`: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22(12):1540–1542

Volkovich Z, Barzily Z, Toledano-Kitai D, Avros R (2010) The Hotteling's metric as a cluster stability index. Comput Model New Technol 14(4):65–72

Wang J (2010) Consistent selection of the number of clusters via cross-validation, "Biometrika", 97:893–904