Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

https://www.czasopisma.uni.lodz.pl/foe/
6(357) 2021
DOI: https://doi.org/10.18778/0208-6018.357.04

Dorota Rozmus https://orcid.org/0000-0002-0565-5319

University of Economics in Katowice, Faculty of Finance, Department of Economic and Financial Analysis, Katowice, Poland, drozmus@ue.katowice.pl

The Number of Groups in an Aggregated Approach in Taxonomy with the Use of Stability Measures and Classical Indices – A Comparative Analysis

Abstract: Recently, the two concepts that have been often discussed in the literature on taxonomy are the cluster ensemble and stability. An interesting proposal regarding the combination of these two concepts was presented by Şenbabaoğlu, Michailidis, and Li, who proposed as a measure of stability a proportion of ambiguously clustered pairs (PAC) for selecting the optimal number of groups in the cluster ensemble. This proposal appeared in the field of genetic research, but as the authors themselves write, the method can be successfully used also in other research areas.

The aim of this paper is to compare the results of indicating the number of clusters (*k* parameter) using the aggregated approach in taxonomy and the above-mentioned measure of stability and classical indices (e.g. Caliński–Harabasz, Dunn, Davies–Bouldin).

Keywords: taxonomy, clustering, cluster ensemble, cluster stability

JEL: C38

1. Introduction

Achieving high accuracy of results is a very important task in any clustering problem. It determines effectiveness of decisions based on research findings. The ability to recognise the actual structure of classes is considered as the clustering accuracy method. Therefore, methods and solutions whose main aim is to give more accurate results than traditional clustering algorithms are proposed in the literature (e.g. *k*-means, *k*-medoids or hierarchical methods). Examples of such solutions can be cluster ensembles (Leisch, 1999; Fred, Jain, 2002; Dudoit, Fridlyand, 2003; Monti et al., 2003; Hornik, 2005; Kuncheva, Vetrov, 2006).

The stability of a taxonomy algorithm against minor changes in a data set (e.g. subtraction from a dataset, small changes in variable values) or algorithm parameters (e.g. random selection of parameter values) is the desired property of the method. In the literature, it is assumed that, with the correct parameters selected, the multiple uses of a given algorithm should give rise to very few differences (i.e. the results should be stable) and reveal the actual structure present in the data. This criterion is particularly applicable when selecting the number of groups (parameter *k*). The literature proposes a number of different ways for measuring stability (e.g.: Ben-Hur, Guyon, 2003; Suzuki, Shimodaira, 2006; Henning, 2007; Brock et al., 2008; Shamir, Tishby, 2008; Volkovich et al., 2010; Fang, Wang, 2012; Lord et al., 2017; Marino, Presti, 2019).

An interesting proposal of a stability measure in a cluster ensemble was presented by Şenbabaoğlu, Michailidis, and Li (2014). Based on the consensus clustering proposed by Monti et al. (2003) and their criteria for selecting the number of groups (i.e. empirical CDF and proportional area change under CDF (ΔK)), Şenbabaoğlu, Michailidis, and Li (2014) introduced the proportion of ambiguously clustered pairs (PAC) as a stability measure.

The rest of the paper is organised as follows. In the next section, the consensus clustering method is described. In the third section, the PAC stability measure is presented. The empirical results are discussed in the fourth section and final conclusions are presented at the end of the paper.

2. Consensus clustering

Consensus clustering is a kind of aggregated approach in taxonomy. The main idea of this approach is to repeatedly cluster a set of data under a certain degree of random perturbation (e.g. resampling) and to calculate the 'consensus index' between all pairs of observations. This index is calculated as the frequency with which a given pair is clustered together into the same group over multiple runs of the clustering algorithm.

The dataset will be denoted as $X = \{x_1, x_2, \dots, x_N\}$ and the first step is to prepare R perturbed datasets $\boldsymbol{X}^r$ (r = 1, 2, ..., R) obtained by resampling the original dataset. Two main concepts in consensus clustering are connectivity matrix and consensus matrix.

Connectivity matrix $\boldsymbol{C}^r$ is an (N x N) matrix created after applying the clustering method to the randomly selected subset of $\boldsymbol{X}^r$. Elements of this matrix are computed as:

$$C^r(i,j) = \begin{cases} 1 & \text{when } i \text{ and } j \text{ item belong to the same cluster;} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

However, it should be noted that the idea of such encoding is not new. In 1976, Sokołowski proposed a similarity measure[1] of the data partitions obtained by different clustering methods, based on the same concept that was later applied by Monti et al. (2003) in the connectivity matrix.

The consensus matrix $\boldsymbol{S}^k$ in clustering into k groups is an (N x N) matrix whose elements indicate the percentage of partitions (among all selected R subsets) in which pairs of observations (i, j) were in the same cluster. The elements of the consensus matrix (marked S (i, j)) are called the consensus index for the appropriate pair of points and are computed as the normalised sum of connectivity matrices over all subsets of $\boldsymbol{X}^r$ (r = 1, 2, ..., R):

$$S(i,j) = \frac{\sum_r C^r(i,j)}{\sum_r I^r(i,j)} . \quad (2)$$

Ir (i, j) is an (N × N) indicator matrix such that its (i, j)-th entry is equal to 1 if both items i and j are present in the dataset $\boldsymbol{X}^r$, and 0 otherwise. The need for the indicator matrix is due to the use of resampling. Some sampling schemes do not include all items from the original dataset.

Each entry in $\boldsymbol{S}^k$ is a real number between 0 and 1. A perfect consensus corresponds to a consensus matrix with all the entries equal to either 0 or 1. This property of the consensus matrix suggests a method for finding the number of clusters that best fits the data. Assuming that a perfect consensus translates into a consensus matrix with

[1] An English-language description of the measure can be found in Sokołowski, 1995: 195–199.

all the entries set to either 1 or 0, it is possible to interpret a deviation from this optimal scenario as an indication of lack of stability of the putative clusters. Hence, based on this property of consensus matrix, Şenbabaoğlu, Michailidis, and Li (2014) proposed the proportion of ambiguously clustered pairs (PAC) as a stability measure of a given cluster solution.

3. PAC stability measure

Given the consensus matrix of the clustering into k groups ($\boldsymbol{S}^k$), a histogram of matrix entries is drawn for $\frac{N \cdot (N-1)}{2}$ consensus indices $S(i, j)$, for $i < j$.

A perfect consensus for pairs of objects among the R partitions will be represented in the histogram by two bars: of zero and of one (Figure 1). The middle part of the histogram (bars between 0 and 1) maps ambiguity clustering for a pair of objects among R partitions (Figure 2).

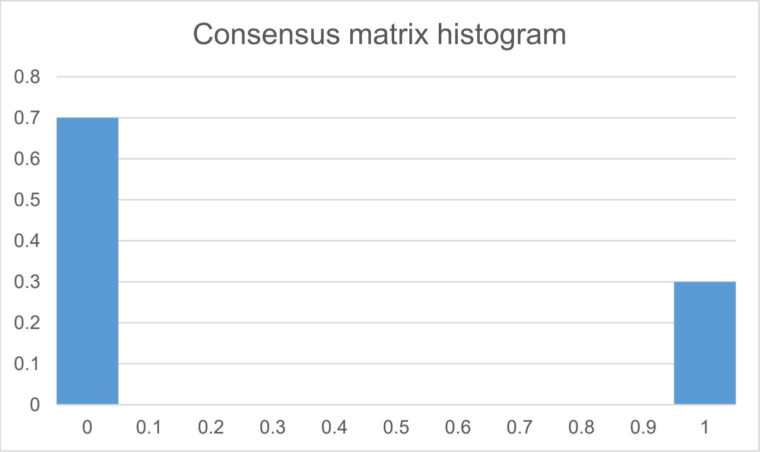

Figure 1. A perfect consensus for pairs of objects among the R partitions

Source: the author's own elaboration

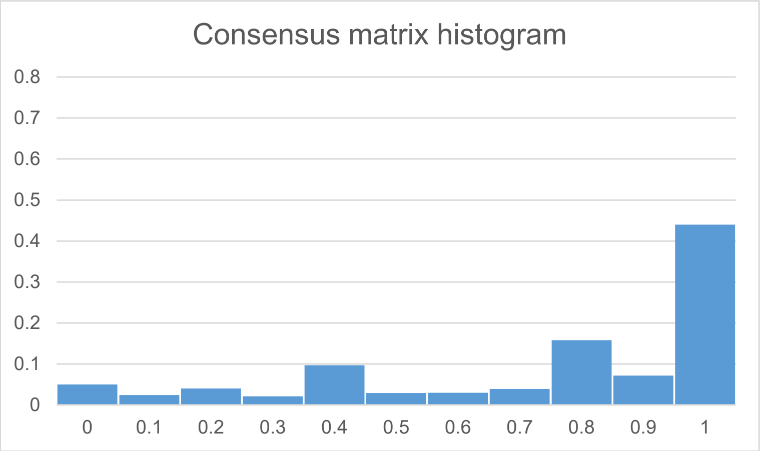

Figure 2. Ambiguous clustering for pairs of objects among *R* partitions
Source: the author's own elaboration

The next step in the construction of a PAC stability measure is to plot the corresponding CDF (empirical cumulative distribution) defined in the range [0, 1]. For a given histogram, it is defined as follows:

$$CDF(w)=\frac{\sum_{i<j} I\{S(i,j)\le w\}}{\frac{N(N-1)}{2}}, \qquad (3)$$

where *I*{.} is an indication function.

Looking at the CDF function, one should note that:

1) the lower left part represents points that never or rarely appear together in the same cluster;
2) the upper right part of the graph represents points that are often or always together in the same cluster;
3) the middle part represents those objects whose coexistence with other objects in the same cluster differs depending on the selected subset of observations.

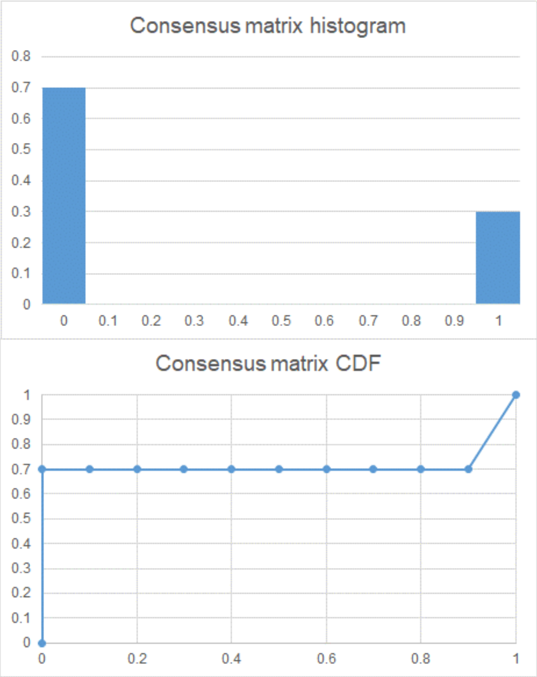

Figure 3. Consensus matrix histogram and corresponding CDF for a perfect consensus for pairs of objects among the *R* partitions

Source: the author's own elaboration

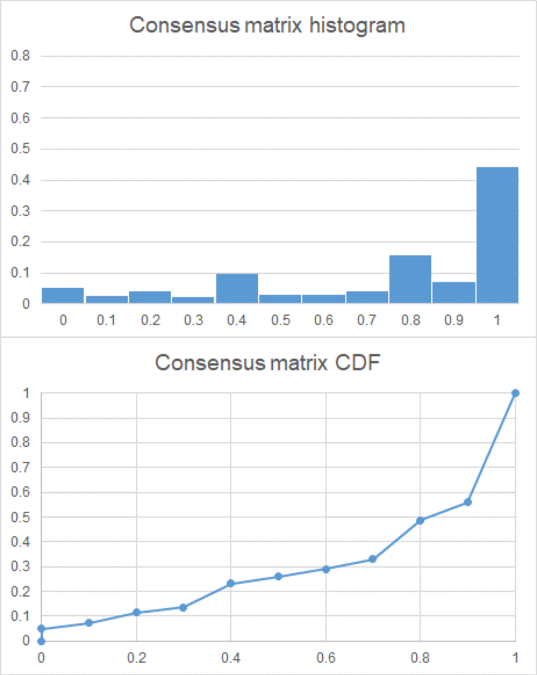

Figure 4. Consensus matrix histogram and corresponding CDF for ambiguous clustering for pairs of objects among the *R* partitions

Source: the author's own elaboration

Therefore, it can be deduced, as it is shown in Figures 3 and 4, that the CDF plot is able to show the optimal number of clusters, because the CDF curve is flat in the middle only for the true value of the parameter *k*.

Using this feature of the CDF plot, the PAC was defined as the fraction of pairs of observations whose consensus index *S* (*i*, *j*) falls into the interval (x_1, x_2) falling within the range [0, 1].

$$PAC_k\left(x_1,x_2\right)=CDF_k\left(x_2\right)-CDF_k\left(x_1\right). \qquad (4)$$

Low PAC values suggest a flat middle part of the CDF plot, and thus the optimal value of *k* is given as:

$$k_{opt}=arg\min_k PAC_k. \qquad (5)$$

4. Empirical results

The goal of empirical experiments is to cluster the European Union countries in terms of sustainable development goals and compare the compliance of indicating the *k* parameter value when changing the base methods of building a cluster ensemble and changing the criteria for selecting the value of the *k* parameter. The data were taken from Eurostat (2019) and refer only to three goals, i.e.:

1) Goal 8 – decent work and economic growth;
2) Goal 9 – industry, innovation and infrastructure;
3) Goal 12 – responsible consumption and production.

All calculations were carried out in the **R** using diceR package (Chiu, Talhouk, 2018).

Among base methods used for consensus clustering construction were: the hierarchical average method (Anderberg, 1973; Gordon, 1987; 1996), *k*-means (Aldenderfer, Blashfield, 1984; Everitt, Landau, Leese, 2001) and *k*-medoids (Kaufman, Rousseeuw, 1990). The considered range for the number of clusters was $k \in (2,3,\ldots,7)$.

For the most important parameters in consensus clustering, i.e. the number of sub-samples (*B*) and the proportion of objects selected for sub-samples, it was assumed: *R* = 10 and 70% of the observations from the original data set for each sub-sample.

When calculating the PAC stability measure, the same values of x_1 and x_2 were assumed as in the experiments of the authors of the method (0.1 and 0.9, respectively).

Among the classical indices used to evaluate the results of the clustering results and to select the *k* parameter, the following indices were used: Calinski–Harabasz (Caliński, Harabasz, 1974; optimisation direction: maximum), Dunn (1974; optimisation direction: maximum) and Davies–Bouldin (Davies, Bouldin, 1979; optimisation direction: minimum).

When analysing the results, two comparative perspectives will be adopted:

1) the first – in which agreement of the criteria under consensus clustering with each base method will be analysed separately (i.e. separately for average, separately for *k*-means and separately for *k*-medoids);
2) the second – in which the agreement of the indices themselves will be assessed, regardless of the base method used in consensus clustering (i.e. separately for PAC, separately for Calinski–Harabasz, separately for Dunn, and separately for Davies–Bouldin).

5. Decent work and economic growth

Looking at the results presented in Table 1, and adopting the first comparative perspective, it can be seen that the only agreement can be observed for the Dunn and Davies–Bouldin index in the *k*-means as the base method for consensus clustering, followed by Calinski–Harabasz and Dunn in the case of *k*-medoids as the base method. For any of the base methods, we find no agreement between classical indices and the PAC stability measure. It is also worth noting that the PAC stability measure suggests very extreme values (*k* = 7 for average and *k* = 2 for *k*-means and *k*-medoids).

Table 1. PAC and indices value for different values of *k* parameter and different base methods in consensus clustering

	Average				*k*-means				*k*-medoids			
k	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**
2	0.323	28.226	0.256	0.463	0.090	43.679	0.035	0.548	0.116	43.679	0.035	0.548
3	0.434	68.063	0.148	0.516	0.212	85.101	0.244	0.457	0.296	84.269	0.084	0.456
4	0.265	77.664	0.108	0.451	0.222	77.664	0.108	0.451	0.238	71.045	0.076	0.407
5	0.161	78.945	0.055	0.500	0.241	72.036	0.080	0.477	0.214	37.168	0.048	0.515
6	0.127	72.110	0.055	0.458	0.177	79.376	0.074	0.540	0.206	81.040	0.055	0.481
7	0.095	38.090	0.038	0.492	0.175	34.344	0.051	0.454	0.127	74.203	0.055	0.441

Source: own computation

Looking at the results presented in Table 1 from the second comparative perspective, it can be concluded that the Davies–Bouldin index suggests the same value of the *k* parameter (*k* = 4), regardless of the chosen base method in consensus clustering. The Calinski–Harabsz index suggests *k* = 3 for the *k*-means and *k*-medoids base methods. For the same base methods, the PAC stability measure is also consistent, though it suggests *k* = 2.

6. Industry, innovations and infrastructure

Based on the results presented in Table 2, and adopting the first comparative perspective, it can be stated that the agreement of the indices under the base methods is only for Calinski–Harabasz, Dunn and Davies–Bouldin for the average base method (indicating *k* = 2) and *k*-medoids (suggesting *k* = 4). For the *k*-means as the base method, only the Dunn and Davies–Bouldin indices agree, suggesting *k* = 2. Again, there is no agreement between the PAC stability measure and any other classical index evaluating clustering results.

Table 2. PAC and indices value for different values of *k* parameter and different base methods in consensus clustering

	Average				*k*-means				*k*-medoids			
k	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davie–Bouldin**	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**
2	0.291	89.058	1.618	0.077	0.373	89.058	1.618	0.077	0.333	27.032	0.088	0.707
3	0.185	49.471	0.119	0.645	0.241	236.485	0.338	0.309	0.222	127.470	0.097	0.405
4	0.212	35.880	0.050	0.604	0.169	251.659	0.141	0.332	0.101	251.659	0.141	0.332
5	0.090	26.798	0.050	0.605	0.140	25.928	0.009	0.603	0.077	26.798	0.050	0.605
6	0.050	10.809	0.021	0.752	0.172	5.213	0.002	1.327	0.143	20.620	0.009	0.604
7	0.098	4.466	0.016	1.213	0.180	4.354	0.002	0.986	0.111	6.275	0.003	0.778

Source: own computation

Looking at the results presented in Table 2 from the second comparative perspective, it can be noted that the index of Dunn and Davies–Bouldin is consistent with the reported value of the number of groups (*k* = 2) for the average and *k*-means base method. Calinski–Harabasz is consistent for *k*-means and *k*-medoids, suggesting *k* = 4. For the same base methods, the PAC stability measure is also consistent, indicating the value of *k* = 3.

Interesting results can be seen in Table 3, where for the average as the base method in clustering consensus construction, the Dunn and Davies–Bouldin indices agree, suggesting clustering into two clusters. The PAC stability measure and the Calinski–Harabasz index also agree in this base method, suggesting $k = 7$ as the correct one. It is worth noting that this is the only case of agreement between the stability measure and any classical index. For the *k*-means and *k*-medoids as the base methods, the same conclusions can be drawn: the Calinski–Harabasz and Davie-Bouldin indices are consistent with each other and suggest $k = 5$.

Table 3. PAC and indices value for different values of *k* parameter and different base methods in consensus clustering

	Average				*k*-means				*k*-medoids			
k	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**	**PAC**	**Calinski–Harabasz**	**Dunn**	**Davies–Bouldin**
2	0.317	8.518	0.560	0.303	0.479	18.860	0.116	1.056	0.394	17.099	0.168	1.058
3	0.437	17.066	0.146	0.814	0.455	18.774	0.130	0.937	0.291	16.238	0.123	1.060
4	0.426	19.931	0.198	0.649	0.317	14.327	0.097	1.116	0.249	14.737	0.123	1.008
5	0.288	9.843	0.116	1.296	0.270	21.514	0.205	0.767	0.241	19.626	0.195	0.819
6	0.230	11.008	0.165	1.191	0.262	20.223	0.206	0.847	0.183	10.267	0.109	1.207
7	0.140	22.419	0.416	0.667	0.185	8.898	0.092	0.985	0.151	9.716	0.092	1.088

Source: own computation

Looking at the results presented in Table 3 from the second comparative perspective, for the first time, regardless of the chosen base method for consensus clustering construction, clearly noticeable is the compatibility of the PAC stability measure, which indicates $k = 7$. The Dunn index is consistent for the average and *k*-medoids as the base method and suggests $k = 2$. The Calinski–Harabasz and Davies–Bouldin indices are consistent for two base methods, i.e. *k*-means and *k*-medoids, indicating $k = 5$ as the number of clusters.

8. Conclusions

The article discusses the problem of choosing the number of clusters (k parameter) in the cluster ensemble using the PAC stability measure and the classical indices, i.e.: the Calinski–Harabasz, Dunn and Davies–Bouldin indices. The PAC measure is a measure of stability dedicated to determining the number of clusters in an aggregated approach which was proposed by Şenbabaoğlu, Michailidis, and Li (2014). This is another concept of the stability measure, which admittedly refers to the cluster ensemble, but the philosophy is still the same: the value of the k parameter indicated by the stability measure should indicate the actual structure of the groups.

Two perspectives were adopted in the comparative study: the first examining the compliance of the criteria for selecting the number of groups for various base methods for the cluster ensemble construction, and the second investigating the compliance of the same criterion for determining the number of clusters, however, with a changing base method for the cluster ensemble construction. The conducted experiments show that: the value of the parameter k indicated by the classical indices completely differs from the values indicated by PAC (regardless of the adopted base method in the cluster ensemble). Classical indices also usually differ with respect to the indicated value of the k parameter. These differences are visible both within the same algorithm chosen as the base method for the construction of the consensus matrix and by comparing the different methods used as the base for the construction of the cluster ensemble.

References

Aldenderfer M.S., Blashfield R.K. (1984), *Cluster analysis*, Sage, Beverly Hills.

Anderberg M.R. (1973), *Cluster analysis for applications*, Academic Press, New York–San Francisco–London.

Ben-Hur A., Guyon I . (2003), *Detecting stable clusters using principal component analysis*, "Methods in Molecular Biology", no. 224, pp. 159–182.

Brock G., Pihur V., Datta S., Datta S. (2008), *clValid: an **R** package for cluster validation*, "Journal of Statistical Software", vol. 25(4), pp. 1–22, https://doi.org/10.18637/jss.v025.i04

Caliński R.B., Harabasz J. (1974), *A dendrite method for cluster analysis*, "Communications in Statistics", vol. 3, pp. 1–27.

Chiu D.S., Talhouk A. (2018), *diceR: an **R** package for class discovery using an ensemble driven approach*, "BMC Bioinformatics", no. 19, 11, https://doi.org/10.1186/s12859-017-1996-y

Davies D.L., Bouldin D.W. (1979), *A Cluster Separation Measure*, "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 1(2), pp. 224–227.

Dudoit S., Fridlyand J. (2003), *Bagging to improve the accuracy of a clustering procedure*, "Bioinformatics", vol. 19(9), pp. 1090–1099.

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

Eurostat (2019), Database, https://ec.europa.eu/eurostat/web/main/data/database (accessed: 20.11.2021).

Everitt B.S., Landau S., Leese M. (2001), *Cluster analysis*, Edward Arnold, London.

Fang Y., Wang J. (2012), *Selection of the number of clusters via the bootstrap method*, "Computational Statistics and Data Analysis", no. 56, pp. 468–477.

Fred A., Jain A.K. (2002), *Data clustering using evidence accumulation*, "Proceedings of the Sixteenth International Conference on Pattern Recognition", pp. 276–280.

Gordon A.D. (1987), *A review of hierarchical classification*, "Journal of the Royal Statistical Society", ser. A, pp. 119–137.

Gordon A.D. (1996), *Hierarchical classification*, [in:] P. Arabie, L.J. Hubert, G. de Soete (eds.), *Clustering and classification*, World Scientific, Singapore, pp. 65–121.

Henning C. (2007), *Cluster-wise assessment of cluster stability*, "Computational Statistics and Data Analysis", no. 52, pp. 258–271.

Hornik K. (2005), *A CLUE for CLUster ensembles*, "Journal of Statistical Software", no. 14, pp. 65–72.

Kaufman L., Rousseeuw P.J. (1990), *Finding groups in data: an introduction to cluster analysis*, Wiley, New York.

Kuncheva L.I., Vetrov D.P. (2006), *Evaluation of stability of k-means cluster ensembles with respect to random initialization*, "IEEE Transactions on Pattern Analysis & Machine Intelligence", vol. 28(11), pp. 1798–1808.

Leisch F. (1999), *Bagged clustering*, "Adaptive Information Systems and Modeling in Economics and Management Science", Working Papers, SFB, no. 51.

Lord E., Willems M., Lapointe F.J., Makarenkov V . (2017), *Using the stability of objects to determine the number of clusters in datasets*, "Information Sciences", no. 393, pp. 29–46.

Marino V., Presti L.L. (2019), *Stay in touch! New insights into end-user attitudes towards engagement platforms*, "Journal of Consumer Marketing", no. 36, pp. 772–783.

Monti S., Tamayo P., Mesirov J., Golub T. (2003), *Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data*, "Machine Learning", no. 52, pp. 91–118.

Şenbabaoğlu Y., Michailidis G., Li J.Z. (2014), *Critical limitations of consensus clustering in class discovery*, "Scientific Reports", no. 4, 6207, https://doi.org/10.1038/srep06207

Shamir O., Tishby N. (2008), *Cluster stability for finite samples*, "Advances in Neural Information Processing Systems", no. 20, pp. 1297–1304.

Sokołowski A. (1995), *Percentage points of the similarity measure for partitions*, "Statistics in Transition", vol. 2(2), pp. 195–199.

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

Volkovich Z., Barzily Z., Toledano-Kitai D., Avros R. (2010), *The Hotteling's metric as a cluster stability index*, "Computer Modelling and New Technologies", vol. 14(4), pp. 65–72.

Wybór liczby grup w podejściu zagregowanym w taksonomii z wykorzystaniem miar stabilności oraz klasycznych indeksów – porównanie wyników

Streszczenie: We współczesnych rozważaniach z dziedziny taksonomii w literaturze często poruszane są dwa pojęcia: podejście zagregowane oraz stabilność metod grupowania. Do tej pory te były one rozważane osobno. Natomiast ciekawą propozycję w zakresie połączenia tych dwóch pojęć przedstawili Y. Şenbabaoğlu, G. Michailidis i J.Z. Li, którzy zasugerowali podejście zagregowane w taksonomii, połączone z zaproponowaną przez siebie miarą stabilności jako kryterium wyboru optymalnej liczby grup (*k*).

Celem artykułu jest porównanie wyników wyboru wartości parametru *k* za pomocą wspomnianej miary stabilności oraz klasycznych indeksów (np. Calińskiego-Harabasza, Dunna).

Słowa kluczowe: taksonomia, klasteryzacja, podejście zagregowane, stabilność metod taksonomicznych

JEL: C38

Received: 2022-01-14; revised: 2022-03-11. Accepted: 2022-06-13