



**Dorota Rozmus**

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

## Using R Packages for Comparison of Cluster Stability

**Abstract:** The stability of clustering methods is the issue that has attracted a considerable amount of attention of researchers in recent years. In this respect, the major question that needs to be answered seems to be to what extent the structure discovered by a particular method is actually present in the data. The literature proposes a number of different ways of measuring stability. The theoretical considerations have 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`, `fpc`, `ClusterStability`, and `pvclust`. Due to the hypothesis that cluster stability can be the answer to the question about the right number of groups in clustering, the main aim of this article is to compare the results of the studies on clustering stability conducted with three R packages, i.e.: `clv`, `clValid`, and `fpc`.

**Keywords:** clustering, taxonomy, stability

**JEL:** C38

## 1. Introduction

The main problem in taxonomy is to determine whether the groups that we received reflect the actual structure of the groups present in the data. This involves the problem of selecting a “clustering model”, e.g.: the number of groups  $k$ , the distance metric, or the control parameters of an algorithm. It is the stability criterion that increasingly gains in popularity in response to these problems.

Informally, this criterion states that if a cluster algorithm is repeatedly used for independent samples (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, Tishby, 2008). Volkovich et al. (2010) even state that the number of groups that maximises the stability of clustering can serve as an estimate of the “true” number of groups.

The main aim of this article is to compare the results of the studies on clustering stability conducted with three **R** packages, i.e.: `clv`, `clValid`, and `fpc`.

## 2. Package `clv`

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 (implemented by the `cls.stab.sim.ind` function) and a measure based on the pattern-wise agreement concept (the `cls.stab.opt.assign` function).

The algorithm of `cls.stab.sim.ind` function can be described in the following steps:

1. Cluster the original data set in order to obtain the reference partition.
2. Select a random sub-sample of observations from the original data set and group the objects from this sub-sample.
3. Calculate the stability between the reference partition and the partition of the sub-sample using the index of similarity between the two partitions (e.g.: the Rand index).
4. Repeat the procedure several times.
5. Repeat the procedure for different values of  $k$  (the number of groups).

The `cls.stab.opt.assign` function is based on the idea of pattern-wise agreement and pattern-wise stability.

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

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

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

Pattern-wise stability is defined as the fraction of sub-sampled partitions where the sub-sampled labelling of pattern  $i$  agrees with that of the reference labelling, by averaging the pattern-wise agreement:

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

where  $N_i$  – the number of sub-samples where pattern  $i$  appears.

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

$$c(j) = \frac{1}{|L_1 = j|} \sum_{i \in (L_1 = j)} n(i). \quad (3)$$

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

$$S_k = \min_j c(j). \quad (4)$$

### 3. Package `clValid`

The package `clValid` contains functions for validating results of clustering analysis in biology. There are three main types of cluster validation measures available: “internal”, “biological” and “stability”.

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 at a time (Brock et al., 2011). 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 normalised 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:

$$APN = \frac{1}{M \cdot N} \sum_{i=1}^N \sum_{j=1}^M \left( 1 - \frac{n(C^{i1} \cap C^{i0})}{n(C^{i0})} \right), \quad (5)$$

where:

$C^{i,0}$  – represents the cluster containing observation  $i$  using the original clustering (based on all available data),

$C^{i,1}$  – represents the cluster containing observation  $i$  where the clustering is based on the data set with  $l$  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.

## 4. Package `fpc`

The package `fpc` includes two functions for measuring stability: `clusterboot` and `nselectboot`. In the experiments, only the `nselectboot` function was used.

The `nselectboot` function is based on the work of Fang and Wang (2012). The authors 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.

The `nselectboot` function 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 optimising 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  $k = 2, \dots, K$ :

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

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

4. Instability of clustering is calculated as:

$$\hat{s}_B = \frac{1}{B} \sum_{b=1}^B d(\Psi_{X_b^{n^*}, k}, \Psi_{\tilde{X}_b^{n^*}, k}) \quad (7)$$

Based on sections 2, 3, 4, one can see that there are a few quite different ways for measuring the stability of clustering algorithm. The main aim of the next part of the article is to carry out numerical experiments in order to verify the hypothesis that the stability criteria could be the answer to the question about the number of groups related to the issue of taxonomy.

## 5. Numerical experiments

In the study, data sets from the UCI Repository, usually used in comparative analyses in taxonomy, were used. Their short characteristics are shown in Table 1.

Table 1. Characteristics of the data sets

Data set	# of observations	# of characteristics	# of groups
UCI segment	2310	19	7
UCI sat	4435	36	7
UCI optdigits	3823	64	10
UCI spect	80	14	2
UCI movement libras	360	90	15

Source: own work on the basis of [https://archive.ics.uci.edu/ml/data sets.html](https://archive.ics.uci.edu/ml/data%20sets.html)

In the experiments, the number of groups that is shown in Table 1 was used as the information of the maximum value of  $k$  (the number of groups). The only exception was the UCI spect data set, where the maximum  $k$  value equalled 5.

Looking at the results (Table 2 – Table 11), we can see that the results are inconclusive.

Table 2. Values of stability measures for UCI sat data for the  $k$ -means method

Method	Number of groups ( $k$ )					
	2	3	4	5	6	7
clv_sim.ind	0.944	<b>0.993</b>	0.921	0.940	0.933	0.940
clv_opt.assign	0.979	<b>0.997</b>	0.760	0.200	0.353	0.299
clValid_APN	<b>0.002</b>	0.009	0.029	0.033	0.059	0.103
fpc	0.074	<b>0.013</b>	0.051	0.046	0.042	0.043

Source: own computation

Table 3. Values of stability measures for UCI sat data for the hierarchical clustering method

Method	Number of groups ( $k$ )					
	2	3	4	5	6	7
clv_sim.ind	0.976	<b>0.978</b>	0.785	0.882	0.918	0.920
clv_opt.assign	<b>0.962</b>	0.199	0.505	0.600	0.576	0.550
clValid_APN	<b>0.020</b>	0.062	0.160	0.095	0.097	0.096
fpc	<b>0.011</b>	0.095	0.098	0.047	0.050	0.048

Source: own computation

In the case of the UCI sat data set for  $k$ -means (Table 2), three methods of measuring stability (i.e. clv\_sim.ind, clv\_opt.assign, and fpc) indicated that the right number of groups is 3, whereas the clValid\_APN criterion pointed to  $k = 2$  as the real number of groups. A similar situation, to a certain extent, applies to the hierarchical clustering method (Table 3): three criteria (i.e. clv\_opt.assign, clValid\_APN, and fpc) indicated  $k = 3$  as the real number of groups, whereas clv\_sim.ind claimed  $k = 2$  as the true number of groups.

Table 4. Values of stability measures for UCI segment data for the  $k$ -means method

Methods	Number of groups ( $k$ )					
	2	3	4	5	6	7
clv_sim.ind	<b>0.993</b>	0.986	0.863	0.869	0.864	0.877
clv_opt.assign	<b>0.995</b>	0.799	0.557	0.566	0.295	0.431
clValid_APN	<b>0.000</b>	0.001	0.001	0.015	0.047	0.077
fpc	<b>0.012</b>	0.043	0.058	0.066	0.063	0.058

Source: own computation

Table 5. Values of stability measures for UCI segment data for the hierarchical clustering method

Method	Number of groups ( $k$ )					
	2	3	4	5	6	7
clv_sim.ind	<b>1</b>	0.999	0.999	0.998	0.996	0.999
clv_opt.assign	<b>1</b>	0.592	0.628	0.686	0.875	0.593
clValid_APN	<b>0</b>	0.001	0.001	0.001	0.027	0.027
fpc	<b>0.012</b>	0.043	0.058	0.066	0.063	0.058

Source: own computation

For the UCI segment data set, in both methods (i.e.  $k$ -means and hierarchical), all the criteria indicated  $k = 2$  as the true number of groups (Table 4 and Table 5).

Table 6. Values of stability measures for UCI optdigits data for the  $k$ -means method

Method	Number of groups ( $k$ )								
	2	3	4	5	6	7	8	9	10
clv_sim.ind	0.699	0.727	0.786	0.899	0.902	0.930	0.947	0.940	<b>0.957</b>
clv_opt.assign	0.589	<b>0.789</b>	0.568	0.741	0.194	0.450	0.348	0.525	0.294
clValid_APN	0.094	0.229	0.208	0.094	<b>0.081</b>	0.091	0.127	0.138	0.098
fpc	0.150	0.123	0.095	0.064	0.046	0.040	0.033	0.028	<b>0.025</b>

Source: own computation

Table 7. Values of stability measures for UCI optdigits data for the hierarchical clustering method

Method	Number of groups ( $k$ )								
	2	3	4	5	6	7	8	9	10
clv_sim.ind	0.778	0.773	0.849	0.928	<b>0.942</b>	0.865	0.909	0.917	0.922
clv_opt.assign	<b>0.564</b>	0.428	0.438	0.483	0.117	0.278	0.003	0.368	0.368
clValid_APN	0.110	0.224	0.085	<b>0.072</b>	0.100	0.104	0.127	0.153	0.156
fpc	<b>0.147</b>	0.186	0.206	0.214	0.222	0.223	0.229	0.229	0.233

Source: own computation

For the UCI optdigits data set and the  $k$ -means method, the results are again inconclusive (Table 6): two criteria indicated  $k = 10$  as the right number of groups, whereas clv\_opt.assign claimed  $k = 3$ , and clValid\_APN pointed to  $k = 6$  as the right number of groups. For the discussed data set and the hierarchical clustering method (Table 7), two criteria (i.e. clv\_opt.assign and fpc) indicated  $k = 2$  as the right number of groups, while clv\_sim.ind pointed to  $k = 6$  groups, and clValid\_APN showed  $k = 5$  as the true number of groups.

Table 8. Values of stability measures for UCI spect data for the  $k$ -means method

Method	Number of groups ( $k$ )			
	2	3	4	4
clv_sim.ind	<b>0.937</b>	0.852	0.769	0.719
clv_opt.assign	<b>0.747</b>	0.179	0.594	0.507
clValid_APN	0.020	<b>0</b>	<b>0</b>	0.018
fpc	<b>0.061</b>	0.117	0.168	0.148

Source: own computation

For the UCI spect data set clustered with the  $k$ -means method (Table 8), clv\_sim ind, clv\_opt.assign and fpc claimed  $k = 2$  as the right number of groups, whereas clValid\_APN indicated  $k = 3$  or  $k = 4$ . For this data set and the hierarchical clustering method (Table 9), two stability criteria stated  $k = 2$  (clv\_opt.assign and fpc), clValid\_APN indicated  $k = 2$  or 3, whereas clv\_sim.ind pointed to  $k = 5$ .

Table 9. Values of stability measures for UCI spect data for the hierarchical clustering method

Method	Number of groups ( $k$ )			
	2	3	4	5
clv_sim.ind	0.977	0.925	0.980	<b>0.986</b>
clv_opt.assign	<b>0.991</b>	0.966	0.710	0.738
clValid_APN	<b>0</b>	<b>0</b>	0.008	0.006
fpc	<b>0.009</b>	0.029	0.039	0.048

Source: own computation

Table 10. Values of stability measures for UCI movement libras data for the  $k$ -means method

Method	Number of groups ( $k$ )						
	2	3	4	5	6	7	8
clv_sim.ind	0.762	0.818	0.856	0.845	0.875	0.903	0.918
clv_opt.assign	<b>0.887</b>	0.835	0.532	0.302	0.648	0.219	0.325
clValid_APN	<b>0.012</b>	0.043	0.069	0.338	0.128	0.156	0.133
fpc	0.153	0.123	0.102	0.082	0.072	0.056	0.051
Method	Number of groups ( $k$ ) – continued						
	9	10	11	12	13	14	15
clv_sim.ind	0.919	0.931	0.932	0.941	0.939	<b>0.945</b>	0.940
clv_opt.assign	0.360	0.205	0.395	0.368	0.170	0.000	0.237
clValid_APN	0.181	0.094	0.057	0.098	0.104	0.094	0.121
fpc	0.047	0.044	0.041	0.037	0.036	0.034	<b>0.033</b>

Source: own computation

Table 11. Values of stability measures for UCI movement libras data for the hierarchical clustering method

Method	Number of groups ( $k$ )						
	2	3	4	5	6	7	8
clv_sim.ind	0.851	0.830	0.805	0.881	0.891	0.877	0.787
clv_opt.assign	<b>0.878</b>	0.657	0.600	0.200	0.200	0.400	0.400
clValid_APN	<b>0.003</b>	0.025	0.096	0.056	0.007	0.005	0.006
fpc	0.137	0.121	0.099	0.097	0.100	0.112	0.118
Method	Number of groups ( $k$ ) – continued						
	9	10	11	12	13	14	15
clv_sim.ind	0.858	0.898	0.929	0.938	0.952	<b>0.959</b>	0.957
clv_opt.assign	0.400	0.600	0.600	0.669	0.479	0.568	0.399
clValid_APN	0.007	0.002	0.023	0.023	0.012	0.012	0.015
fpc	0.122	0.102	0.076	0.067	0.049	0.044	<b>0.041</b>

Source: own computation



For the UCI movement libras and both clustering methods, the results are the same (Table 10 and Table 11): `clv_opt.assign` and `clValid_APN` pointed to  $k = 2$  as the right number of clusters, while `clv_sim.ind` pointed to  $k = 14$ , and `fpc` indicated  $k = 15$ .

## 6. Conclusions

The stability criterion is becoming an increasingly popular method for the selection of parameters of clustering methods, especially for determining the number of groups  $k$ .

If the taxonomy method is selected correctly and the parameters of this method are also selected correctly (e.g.: the number of groups, the distance metric), then clustering should provide results that are not very different from each other, i.e. the results should be stable.

The empirical results show that the examined stability criteria do not always lead to clear results, providing different answers to the question about the right number of groups in the data.

The methods presented in this article are just some proposed ways for measurement of stability, but not the only ones that can be found in the literature. There are other new methods proposed which can be found, for example, in the works of: Granichin et al. (2015), Hosein et al. (2011), Koepke, Clarke (2013) and Ryazanov (2016).

## References

- Ben-Hur A., Guyon I. (2003), *Detecting Stable Clusters Using Principal Component Analysis*, “Methods in Molecular Biology”, vol. 224, pp. 59–182.
- Brock G., Pihur V., Datta S., Datta S. (2011), *clValid: An R Package for Cluster Validation*, <http://cran.us.r-project.org/web/packages/clValid/vignettes/clValid.pdf>.
- Fang Y., Wang J. (2012), *Selection of the Number of Clusters via the Bootstrap Method*, “Computational Statistics and Data Analysis”, vol. 56, pp. 468–477.
- Granichin O., Volkovich Z., Toledano-Kitai D. (2015), *Cluster Validation*, “Randomized Algorithms in Automatic Control and Data Mining”, vol. 67, pp. 163–228.
- Hosein A., Behrouz M., Hamid P., Mohsen M. (2011), *An Asymmetric Criterion for Cluster Validation*, “Developing Concepts in Applied Intelligence”, Studies in Computational Intelligence”, vol. 363, pp. 1–14.
- Koepke H., Clarke B. (2013), *A Bayesian Criterion for Cluster Stability*, “Statistical Analysis and Data Mining: The ASA Data Science Journal”, vol. 6, issue 4, pp. 346–374.
- Ryazanov V. (2016), *About Estimation of Quality of Clustering Results via Its Stability*, “Intelligent Data Analysis”, vol. 20(1), pp. 5–15.
- Shamir O., Tishby N. (2008), *Cluster Stability for Finite Samples*, “Advances in Neural Information Processing Systems”, vol. 20, pp. 1297–1304.


- Volkovich Z., Barzily Z., Toledano-Kitai D., Avros R. (2010), *The Hotteling's Metric as a Cluster Stability Measure*, "Computer Modelling and New Technologies", vol. 14, no. 4, pp. 65–72.
- Wang J. (2010), *Consistent Selection of the Number of Clusters via Cross-validation*, "Biometrika", vol. 97, pp. 893–904.

### Zastosowanie pakietów programu R do porównania stabilności grupowania

**Streszczenie:** W ostatnich latach dużo uwagi poświęca się zagadnieniu stabilności metod taksonomicznych, czyli odpowiedzi na pytanie o to, na ile struktura odkryta przez daną metodę rzeczywiście jest obecna w danych. W literaturze zaproponowano wiele różnych sposobów pomiaru stabilności. W ślad za rozważaniami teoretycznymi w tym zakresie idzie także rozwój narzędzi informatycznych pozwalających na praktyczne zastosowanie zaproponowanych sposobów badania stabilności. Wśród tych narzędzi jest także kilka bibliotek w programie **R**, np. `clValid`, `clv`, `fpc`, `ClusterStability`, `pvcLust`. Celem artykułu jest porównanie wyników badania stabilności grupowania za pomocą wybranych bibliotek w programie R.

**Słowa kluczowe:** grupowanie, taksonomia, stabilność

**JEL:** C38

	<p>© by the author, licensee Łódź University – Łódź University Press, Łódź, Poland.          This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license CC-BY          (<a href="http://creativecommons.org/licenses/by/3.0/">http://creativecommons.org/licenses/by/3.0/</a>)</p>
<p>Received: 2016-12-16; verified: 2017-05-17. Accepted: 2017-09-07</p>	