Available online at http://www.journalijdr.com

IJDR

International Journal of
DEVELOPMENT RESEARCH

ISSN: 2230-9926

International Journal of Development Research
Vol. 06, Issue, 07, pp.8640-8646, July, 2016

Full Length Research Article

QUALITY EVALUATION OF A SELECTED PARTITION: AN APPROACH BASED ON RESAMPLING METHODS

[1, *]Osvaldo Silva, [2]Helena Bacelar-Nicolau, [3]Fernando Nicolau and [4]Áurea Sousa

[1] University of Azores, Department of Mathematics, CICS.UAc/CICS.NOVA.UAc, Portugal
[2]University of Lisbon, Faculty of Psychology and ISAMB-FML, Portugal
[3]New University of Lisbon, FCT, Department of Mathematics, Portugal
[4]University of Azores, Department of Mathematics, CEEAplA, Portugal

ARTICLE INFO

Article History:
Received 24th April, 2016
Received in revised form
26th May, 2016
Accepted 29th June, 2016
Published online 31st July, 2016

Key Words:
Comparing partitions,
Cluster stability,
Hierarchical cluster analysis,
Resampling methods.

ABSTRACT

The aim of this work on cluster analysis is to provide a methodology to analyse and assess the quality of a selected partition (the best partition according to several validation indexes). In the proposed approach, the evaluation of the stability and of the consistency of the results of the selected partition (original partition) was done using the comparison between this partition and each of the partitions (with the same number of clusters that the original one) obtained by resampling. A special emphasis is given to an index defined by linear combination of four indicators, which allows evaluating the adjustment between the original partition and each of the partitions (and / or set of obtained partitions) obtained from resampling data. The application of these indexes is exemplified using a set of real data, and the main conclusions are summarized and discussed.

INTRODUCTION

Cluster analysis involves the organization of a set of elements into groups (classes or *clusters)* with a high intra-group homogeneity and a high inter-group heterogeneity (Jain and Dubes, 1988). The quality of the results is often difficult to assess, given the existence of a number of factors that influence them and the multiplicity of possible clustering results. Depending on the comparison measures (between elements and between clusters) and of all strategies adopted, different clustering structures can be obtained. It should be noted also that many of the comparison measures between elements are affected by the addition of outliers, which affects the quality of the results. Thus, it is important the use and development of validation methodologies, based on the comparison of partitions using resampling methods. A major reason for using validation, in the context of comparing a set of partitions, is that results of different clustering methods from the same dataset may differ substantially.

An obtained partition is stable when it is not much affected by small changes (either using subsamples, introduction of noise, outliers or missing values) in the initial data (Ben-Hur *et al.,* 2002) or by using different clustering algorithms. Thus, a result of a classification is considered stable when it captures the underlying structure of the dataset under the assumption that this partition can be reproduced with other data obtained from the same dataset (Lange *et al.,* 2004). A low variability in the partitions is interpreted as a high consistency of the obtained results (Cheng and Milligan, 1996). In this context, we highlight the importance of verifying the permanence of the elements to be classified in certain clusters in the set of the partitions to be compared. Several indexes based on the concept of stability have been proposed (e.g., Domany and Levine, 2001; Ben-Hur *et al.,* 2002; Lange *et al.,* 2004; Smolkin and Ghosh, 2003; Fridlyand and Dudoit, 2001; Bertoni and Valentini, 2007). However, it is often unclear how these indexes are used in practice (Hennig, 2004), given that a single index provides a limited classification evaluation. Silva *et al.* (2012) presented a global approach concerning to the evaluation of the quality of clustering results obtained by different clustering algorithms using multiple information

***Corresponding author: Osvaldo Silva,**
Department of Mathematics, University of Azores, CICS.UAc/CICS.NOVA.UAc, Portugal.

(e.g., stability, homogeneity and isolation of the clusters). The purpose of this paper is to analyse some indexes to evaluate the quality of partitions based on several types of information, including the values (corresponding to each level of the dendrograms) concerning with the: *i)* number of common elements in the clusters of the two partitions, *ii)* information associated with the isolation and homogeneity of each of the clusters, *iii)* global statistics of levels, STAT, and *iv)* fusion coefficient, *CF*. The use of these indexes based on a real dataset is illustrated in the section concerning to the main results.

Measures of agreement and stability of the clustering results by resampling

Let E= $\{x_1,..., x_m\}$ be a set of m elements (h =1,..., m), to be classified, $P_o = \{c_1, c_2, ..., c_k\}$a partition into k clusters obtained from the initial data matrix, $P_\ell = \{c_1^{(\ell)}, c_2^{(\ell)}, ..., c_k^{(\ell)}\}$another partition into k clusters, obtained from the data matrix corresponding to the resampling ℓ ($1 \leq \ell \leq r$), and r the number of resamples. Based on this notation, some measures of agreement and stability are referred in this section. The first are related to the two partitions, P_0 and P_ℓ. For each element, h, (h=1,…, m) we compare the cluster in which $h$is included in the original partition (base), P_0, with the cluster in which h belongs in the partition P_ℓ, considering also the information regarding the number of elements in each cluster. The stability indexes are obtained based on the comparison of the original partition, P_0, with each one of the partitions obtained in the r random resamples.

Indexes based only on the common elements

Let c_i (i=1,..,k) be the cluster of partition P_o, obtained from the initial data matrix, $c_i^{(\ell)}$ the corresponding cluster of partition P_ℓ, obtained with the data from resampling ℓ . The agreement index (*AI*) between each cluster c_i, i =1,…,k, of the partition P_o and the corresponding cluster, $c_i^{(\ell)}$,of partition P_ℓ,may be defined by:

$$AI\left(c_i, c_i^{(\ell)}\right) = \max_{1 \leq j \leq k} \left\{ \frac{card\left(c_i \cap c_j^{(\ell)}\right)}{card\left(c_i\right)} \right\}, \qquad (1)$$

The corresponding agreement index between the partitions P_o and P_ℓ may be defined by:

$$AI\left(P_0, P_\ell\right) = \underset{i=1,...k}{Median} \left\{ AI\left(c_i, c_i^{(\ell)}\right) \right\}. \qquad (2)$$

The higher the values obtained through formulae (1) and (2) the better the consistency of the obtained results. The stability index for the cluster c_i belonging to the initial partition P_o may be obtained based on the comparison between this partition, P_o, and the corresponding partitions of the r resamples, according to the following formula:

$$E\left(c_i\right) = \underset{\ell=1,...,r}{Median} \left\{ AI\left(c_i, c_i^{(\ell)}\right) \right\} \qquad (3)$$

The stability index concerning to the original partition, P_o, can be calculated through:

$$E\left(P_0\right) = \underset{\ell=1,...,r}{Median} \left\{ AI\left(P_0, P_\ell\right) \right\}, \qquad (4)$$

Where $AI\left(P_0, P_\ell\right)$ is obtained by formula (2). The higher the values obtained through formulas (3) and (4), the higher will be the stability of the obtained results.

Index based on the concepts of homogeneity and isolation

In cluster analysis, the concepts of homogeneity (compactness) and isolation or separation are behind the idea of classes (clusters). That is, it is intended that a cluster is a set of similar entities and that entities belonging to different clusters are not similar (Everitt, 1993). Based on these concepts, the Silhouette index (Rousseeuw, 1987), known as *Sil*, was developed in order to assess the relative compactness and isolation of clusters, and distinguishing between their core and outlying members. The values of *Sil* are comprised between -1 and +1. Negative values of this index indicate that the element is more similar to elements of another cluster, and values near +1 indicate that the element strongly belongs to the cluster in which it has been placed. The indexes in this section are based on a modified version of this index (Sousa *et al.,* 2014). In this section, we present an agreement index between the original partition, P_0, and the partition obtained through resampling, P_ℓ, based on information associated to the isolation and homogeneity of each cluster ($c_i \cap c_i^{(\ell)}$), with ℓ =1,…,r, where $c_i^{(\ell)}$ is the cluster of the resample ℓ for which a maximum of $card\left(c_i \cap c_j^{(\ell)}\right)$, j=1,…,k, is verified. Therefore, the following steps consist in the calculation of the values of:

i. *Sil* for each element h, $h \in \left(c_i \cap c_i^{(\ell)}\right)$,denoted by *Sil*($h$), and for each cluster$\left(c_i \cap c_i^{(\ell)}\right)$, denoted by $Sil\ classe\left(c_i \cap c_i^{(\ell)}\right)$, based on the proximity matrix concerning to the original data;

ii. *Sil* for each element h, $h \in \left(c_i^{(\ell)} \cap c_i\right)$, denoted by $Sil^{(\ell)}(h)$, and for each cluster$\left(c_i^{(\ell)} \cap c_i\right)$, denoted by $Sil_{(\ell)}classe\left(c_i^{(\ell)} \cap c_i\right)$, based on the proximity matrix related to the data matrix obtained from the ℓ resample ℓ *=1,..., r*;

iii. agreement index of element h, $h \in \left(c_i \cap c_i^{(\ell)}\right)$, using the expression:

$$AI_{Sil}{}^{(\ell)}(h) = \frac{Sil^{(\ell)}(h)}{Sil\ (h)}, \qquad (5)$$

Where *Sil(h)* and $Sil^{(\ell)}(h)$ were indicated, respectively, in *i)* and *ii)*;

iv) agreement index between c_i and $c_i^{(\ell)}$, considering the homogeneity and the isolation of the elements, defined by:

$$AI_{Sil}\left(c_i, c_i^{(\ell)}\right) = \frac{Sil_{(\ell)}classe\left(c_i^{(\ell)} \cap c_i\right)}{Sil\ classe\left(c_i \cap c_i^{(\ell)}\right)}; \qquad (6)$$

v) agreement index between P_0 and P_ℓ, given by:

$$AI_{Sil}\left(P_0, P_\ell\right) = \underset{i=1,\ldots,k}{Median}\left\{AI_{Sil}\left(c_i, c_i^{(\ell)}\right)\right\}; \qquad (7)$$

vi) Stability index of element *h*, $h \in \left(c_i \cap c_j^{(\ell)}\right)$, which can be defined by:

$$E_{Sil}\,(h) = \underset{\ell=1,\ldots,r}{Median}\ \left\{AI_{Sil}{}^{(\ell)}(h)\right\}; \qquad (8)$$

vii) stability index of c_i, defined by:

$$E_{Sil}\left(c_i\right) = \underset{\ell=1,\ldots,r}{Median}\ \left\{AI_{Sil}\left(c_i, c_i^{(\ell)}\right)\right\}; \qquad (9)$$

viii) stability index of P_0, defined by the expression:

$$E_{Sil}(P_0) = \underset{\ell=1,\ldots,r}{Median}\ \left\{AI_{Sil}\left(P_0, P_\ell\right)\right\}; \qquad (10)$$

xi) global consistency index of resample ℓ, defined by:

$$ICG_{Sil}(\ell) = \underset{h=1,\ldots,m}{Median}\left\{AI_{Sil}{}^{(\ell)}(h)\right\} \qquad (11)$$

If the values of the agreement indexes given by formulae (6) and (7) are within the interval [0.975, 1.025], then we must admit that there are similar results between the original partition, *P₀*, and the corresponding partition of the ℓ resample, P_ℓ, as far as homogeneity and isolation are concerned. Thus, if the values in these indexes are within the interval [0.975, 1.025], for the partitions obtained through multiple resamples there will be a greater confidence in the choice of the original partition and in the robustness of the results. In that case, the values obtained through formulas (9) to (10) will be very close to 1. In contrast, values of the agreement index corresponding to formulas (6) and (7) that are less than 0.975 (respectively, higher than 1.025) mean that the corresponding structure of the original partition is better (respectively, worse) than the one of the resample ℓ, regarding to homogeneity and isolation.

Indexes based on the preordination associated with each of the partitions

The global statistics of levels, *STAT,* is a global statistic that measures the information given by the corresponding partition, compared to the initial preordination associated with the used (dis)similarity coefficient (e.g., Lerman 1981; Bacelar-Nicolau, 1980, 1988). A good cut-off level corresponds to a partition in which the STAT index shows a significant increase, relatively to the values provided by the neighboring levels. Once the most significant partition (the best partition according to some validation indexes) of the original dataset, $P_o = \{c_1, c_2, \ldots, c_k\}$, containing *k* clusters is chosen, we will compare it with the partitions with the same number of clusters obtained through resampling, noting the number of common elements in each of the *k* clusters of the two partitions and identifying the set of pairs assembled in the same cluster and the set of separated pairs. Subsequently, the values of the agreement and stability indexes defined by the following expressions can be calculated:

$$AI_{STAT}\left(\mathrm{P}_0, \mathrm{P}_\ell\right) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} C_2^{card\left(\mathrm{c_i} \cap \mathrm{c_i^{(\ell)}}\right)}}{C_2^{\mathrm{n}}} \times \frac{STAT_{\mathrm{P}_\ell}}{STAT_{\mathrm{P}_0}} \qquad (12)$$

And

$$E_{STAT}\left(P_0\right) = \underset{\ell=1,\ldots,r}{Median}\ \left\{AI_{STAT}\left(P_0, P_\ell\right)\right\} \qquad (13)$$

If preordination on a certain level, *v*, is equal for the two partitions being compared, the respective *STAT* values are equal and hence their ratio, which is part of formula (12) is equal to the unity. The weight used in determining the index $AI_{STAT}\left(P_0, P_\ell\right)$ takes into account the number of pairs of elements together in the same cluster in the partitions P_0 and P_ℓ in relation to the total number of pairs. The higher the value of $AI_{STAT}\left(P_0, P_\ell\right)$ that is, the closer of the unity, the greater the degree of agreement between the two partitions. Thus, the value of $E_{STAT}\left(P_0\right)$ gives us an indication of the degree of stability of the original partition.

Indexes based on the fusion coefficient

The fusion coefficient is the numerical value at which various cases merge to form a cluster in the context of hierarchical methods. For each of the obtained dendrograms, the corresponding fusion coefficient values, *CF*, may be noted for each of the levels, *v*, then determining the rate of variation, ΔCF, of the values of this coefficient when switching from level *v*-1 to level *v,* by the following expression:

$$\Delta CF(v) = \frac{CF(v) - CF(v-1)}{CF(v-1)}, \qquad (14)$$

With *v*=1,…., nivmax, where nivmax is the maximal number of levels of the dendrogram. As is referred in Aldenderfer and Blashfield (1984), a jump implies that two relatively dissimilar clusters have been merged. Therefore, the number of clusters prior to the merger corresponds to the best cut-off. Mojena (1977) and Mojena and Wishart (1980) have attempted to define quantitative criteria for identifying a “significant jump” in the values of the fusion coefficients. The index given by formula (14) is a reference indicator to choose the appropriate number of clusters. Note that the fusion coefficient depends on the measures of comparison between elements and between clusters. Thus, if it is desired to compare dendrograms obtained from different (dis)similarity matrices and / or different aggregation criteria, a standardization of the values of the fusion coefficients during the tree construction is recommended, so that their values vary between 0 and 1. Thus, for each dendrogram, the values of the fusion coefficient CF *(v), v* = 1, …., *nivmax can* be divided, for example, by the maximum of these values, CF_{max}, which is obtained in the case in which the elements are all assembled in the same cluster, thereby obtaining standardized values which are here denoted by $\phi_{CF}(v)$. Let $\phi_{CF(k)}^{(\ell)}$ and $\phi_{CF(k)}^{o}$ be values obtained in the partition level corresponding to *k* clusters, respectively, in resample ℓ , with ℓ=1,…,*r*, and the original classification. Using this notation, we can calculate the ratio between the values of $\phi_{CF(k)}^{o}$ from the original partition, P_0, and $\phi_{CF(k)}^{(\ell)}$, from the partition P_ℓ, with ℓ=1,…,*r*, defined by the expression:

$$AI_{CF}(P_0,P_\ell)=\frac{\phi_{CF(k)}^{(\ell)}}{\phi_{CF(k)}^{o}}. \qquad (15)$$

A value of $AI_{CF}(P_0,P_\ell)$close to 1 indicates a strong association between the original partition and the partition obtained in there sample ℓ. Note, in particular, that if the condition $AI_{CF}(P_0,P_\ell)\cong 1, \forall \ell \in \{1,\cdots,r\}$ is verified, the original partition, P_0 will present a very high stability. We can also calculate the stability index of partition P_0, using the following formula:

$$E_{CF}(P_0)=\underset{\ell=1,\ldots,r}{Median}\ \{AI_{CF}(P_0,P_\ell)\}, \qquad (16)$$

Where values close to 1 indicate a strong stability of the original partition. Note that, the Sil, STAT and CF indexes are based in different types of information and their values depend on the used comparison measures. Therefore, what we propose are modified functions of these indexes in order to standardize their values. The $AI(P_0 ,P_\ell)$, $ICG_{Sil}(\ell)$, $AI_{STAT}(P_0,P_\ell)$, and $AI_{CF}(P_0,P_\ell)$ indexes given by the formulae (2), (11), (12), and (15) may be combined in a table, such as Table 1. In addition, based on this table, we can assess the degree of agreement between the four indicators in the set of *r* resamples using the Kendall correlation coefficient, in order to evaluate the agreement between these indicators in the overall set of resamples. Based on the linear combination of the indexes presented in Table 1, we consider the index defined bythe following expression:

$$Ajust(\ell)=\alpha_1{*}\ AI(P_0 ,P_\ell)+\alpha_2{*}\ ICG_{Sil}(\ell)+\alpha_3{*}\ AI_{STAT}(P_0,P_\ell)+\alpha_4{*}\ AI_{CF}(P_0,P_\ell), \qquad (17)$$

where α_1, α_2, α_3 and α_4 are the weights associated with each of the coefficients, $\alpha_1+\alpha_2+\alpha_3+\alpha_4=1$, and ℓ =1,…,*r*. The *Ajust* (ℓ) index aims to assess the adjustment between the original partition and each of the partitions referring to each one of the resamples. The global adjustment between the original partition and the set of partitions obtained through resampling may be evaluated based on the values of the index Ajust (ℓ) in the set of the *r* resamples, as shown in the following expression:

$$Ajust\_Glob=\underset{\ell=1,\ldots,r}{Median}\ \{Ajust(\ell)\}\cdot \qquad (18)$$

In case we want to compare various methods, the values obtained by the *Ajust_Glob* index for each of the methods allow their sorting according to the obtained degree of adjustment. Thus, we obtain an indication of which methods provide a more similar structure to that of the partition considered to be the most significant in the original data set. Considering also the values of the indexes $AI(P_0 ,P_\ell)$, $ICG_{Sil}(\ell)$, $AI_{STAT}(P_0,P_\ell)$, and $AI_{CF}(P_0,P_\ell)$ obtained in the *r* resample, as shown in Table 1, one can resort to Kendall concordance coefficient and the corresponding significance test, for an overall assessment of the degree of agreement of the *r* resample in relation to the values of these indexes.

RESULTS

The Ascending Hierarchical Cluster Analysis (AHCA) of twenty-two variables (items) present in a questionnaire, which correspond to statements concerning attitudes / beliefs of the students towards Statistics in the courses of Human and Social Sciences in Higher Education (Silva *et al.,* 2007, 2009), was performed in order to obtain a typology of the variables. The AHCA was based on the affinity coefficient combined with twenty five aggregation criteria, twelve of which were classical and thirteen probabilistic (Nicolau, 1983; Nicolau and Bacelar-Nicolau, 1998).In the case of the analyzed dataset, using only the information given by the data themselves and the set of partitions obtained from twenty-five aggregation criteria, the selected partition, according to the γ (Goodman and Kruskal, 1954), DIF (Bacelar-Nicolau, 1980), Sil (Sousa *et al.,* 2014) indexes and to the U Mann Whitney statistics (1947), is a partition into four clusters: c_1: {*V1, V2, V11, V14, V15*}*;* c_2: {*V3, V9, V21, V22, V4, V16, V20, V6,* V17, V19}; c_3: {*V5, V7, V8*}; and c_4: {*V10, V13, V18, V12*}. In order to undertake an assessment of the quality of this partition, we opted for the use of resampling within a methodology for evaluating and comparing partitions (Silva, 2011). Fifty sub-samples were obtained from the initial data (sampling rate of 80%) using the simple random sampling. Then, we applied the same AHCA algorithm to data matrices corresponding to the various sub-samples and wrote the constitution of the partitions into four clusters for each one of the resamples. In the context of comparing the most significant partition obtained from the initial data set with each partition obtained from the fifty subsamples, we used the indexes presented in the second section for the case of usage of the resampling methods, so as to test the stability and agreement of the obtained results. In Table 2 we present the values relative to the indexes $AI(c_i,c_i^\ell)$ and $AI(P_0 ,P_\ell)$, with *i*=1,…, 4 and ℓ =1,…,50, which correspond, respectively, to the formulae (1) and (2). We also present the values of the $E(c_i)$and $E(P_0)$ indexes to assess, respectively, the stability of each cluster and of the original partition. In the case of $AI(P_0 ,P_\ell)$, $E(c_i)$ and $E(P_0)$, the first value of each one of these indexes, in Table 2, corresponds to the application of formulae (2), (3) and (4), respectively, and the second one (separate by semicolon) corresponds to the use of the mean instead of the median in these formulae.

Table 1 – Values of $AI(P_0, P_\ell)$, $ICG_{SIL}(\ell)$, $AI_{STAT}(P_0, P_\ell)$ and $AI_{CF}(P_0, P_\ell)$ in r resamples

Resample ℓ	$AI(P_0, P_\ell)$	$ICG_{SIL}(\ell)$	$AI_{STAT}(P_0, P_\ell)$	$AI_{CF}(P_0, P_\ell)$
1	$AI(P_0, P_1)$	$ICG_{SIL}(1)$	$AI_{STAT}(P_0, P_1)$	$AI_{CF}(P_0, P_1)$
2	$AI(P_0, P_2)$	$ICG_{SIL}(2)$	$AI_{STAT}(P_0, P_2)$	$AI_{CF}(P_0, P_2)$
…	…	…	…	…
r	$AI(P_0, P_r)$	$ICG_{SIL}(r)$	$AI_{STAT}(P_0, P_r)$	$AI_{CF}(P_0, P_r)$

Table 2. Evaluation of agreement and stability of the clusters and of the original partition

			Agreement		
Resample ℓ	$AI(c_1, c_1^\ell)$	$AI(c_2, c_2^\ell)$	$AI(c_3, c_3^\ell)$	$AI(c_4, c_4^\ell)$	$AI(P_0, P_\ell)$
1	0.8	0.5	1	0	0.65; 0.575
…	…	…	…	…	…
50	1	0,9	1	1	1; 0.975
			Stability		
	$E(c_1)$	$E(c_2)$	$E(c_3)$	$E(c_4)$	$E(P_0)$
	1; 0.988	1; 0.976	1; 0.98	1; 0.88	1; 0.956

Table 3. Agreement and stability indexes based on the *Sil* values

h	$AI_{Sil}^{(1)}(h)$	$AI_{Sil}^{(2)}(h)$	…	$AI_{Sil}^{(49)}(h)$	$AI_{Sil}^{(50)}(h)$	$E_{Sil}(h)$
1	0.999762	0.999359	…	1.002490	1.000806	1.00
15	0.000000	0.000000	…	0.000000	0.999416	0.89
…	…	…	…	…	…	...
8	0.994801	0.998483	…	1.000547	1.001168	1.00
12	0.000000	0.998294	…	1.001590	0.998015	0.78
$ICG_{Sil}^{(\ell)}$	0.54	0.95	…	0.96	0.95	

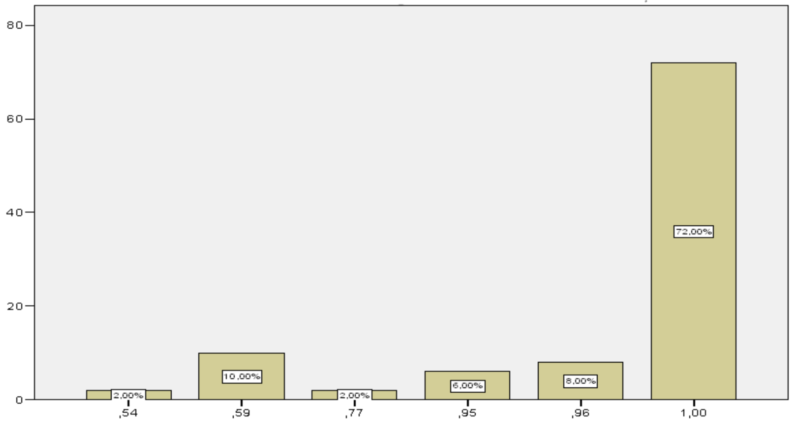

Figure 1. Frequency distribution (%) of the values of $ICG_{SIL}(\ell)$, ℓ =1,…, 50

The large values of the median concerning to the $AI(P_0, P_\ell)$, $E(c_i)$ and $E(P_0)$ indexes showed in Table 2 point to a good agreement and stability of the clusters and of the original partition (most values are equal to one). Moreover, the corresponding mean values allow to complement the information associated with these indexes in order to identify, for example, the more stable clusters (in this case are the first three clusters). Thus, in Figure 1, we present the values of the ICG$_{SIL}$ (ℓ) index using the mean instead of the median in the formula (11). Table 3 shows the results obtained from the application of the $AI_{Sil}^{(\ell)}(h)$, $E_{Sil}(h)$, and $ICG_{Sil}^{(\ell)}$ indexes. From this table the values may be obtained from other indexes, namely from the $E_{Sil}(c_i)$ index, *i*=1,…,4, which values are, respectively, 0.99, 0.93, 0.98, and 0.88, and from the $E_{Sil}(P_0)$ index which value is 0.94. As can be seen from Table 3, the values of $E_{Sil}(h)$ index, *h*=1,…,22, defined by formula (8), support the conclusion that the variables *V1, V11, V2, V14, V9, V21, V22, V7*, and *V8* have a perfect fit, while the variables *V4, V16, V20, V6*, and *V17* are those which have a lower degree of stability due to the presence of these variables not in the same cluster in some of the partitions obtained by resampling. It appears that the most stable clusters are c_1 and c_3, while the less stable is the cluster c_4, as indicated by $E_{Sil}(c_i)$ index, *i*=1,…,4. It is noteworthy the high stability of the original partition, as can be concluded from the value (0.94) obtained by the $E_{Sil}(P_0)$ index. From Figure 1, it is concluded that, although there is an appreciable amplitude with respect to the values of the global consistency index, $\mathrm{ICG}_{\mathrm{SIL}}(\ell)$, $\ell=1,\ldots,50$, each of the r resamples (formula (11)) when compared with the original partition, most of these resamples (72%) have a perfect consistency, with only a small fraction of resamples (2%) having a relatively low consistency. The Ajust (ℓ) index given by formula (17) returned the highest values concerning resample partitions that had a more similar structure to that of the original partition. In this example, we included in formula (17) only the $AI\left(P_0, P_\ell\right)$ and $ICG_{Sil}(\ell)$ indexes. The obtained value (0.908) for the *Ajust_Glob* index corresponding to formula (18) is an indication that the original partition, P_0, is robust.

DISCUSSION

The indexes presented in this paper allow us to assess the adjustment between the original partition and each of the partitions obtained from resampling methods, using the multiple information provided by some agreement indexes, calculated for the several levels of the dendrograms. We have also presented a general adjustment (*Ajust_Glob*) index. The values associated with the Ajust (ℓ) index in the various resamples were calculated in order to evaluate the global adjustment between the original partition and each of the partitions belonging to each one of resamples, while the values obtained by the *Ajust_Glob* index aim to assess the overall adjustment between the original partition and all the partitions obtained by resampling. In the case of the data set "Attitudes and Beliefs towards Statistics" the application of the used indexes allowed us to verify that the partition under evaluation is robust and consistent. The used indexes allow a better evaluation of the stability and consistency of a selected partition, since a high stability is associated with a high consistency of the results. However, it is important to note that obtaining stable results does not necessarily imply the identification of the most concordant partition with the actual existing structure in the population. The use of various strategies to evaluate the agreement and stability of the reference partition gives us the possibility to obtain additional information we consider relevant to support the situation under analysis.

REFERENCES

Aldenderfer, M.S. and Blashfield, R.K. 1984. Cluster Analysis. Sage University Papers.

Bacelar-Nicolau, H. 1980. Contributions to the study of comparison coefficients in cluster analysis.Ph.D.Thesis (in Portuguese), Universidade de Lisboa, Portugal.

Bacelar-Nicolau, H. 1988. Two Probabilistic Models for Classification of Variables in Frequency Tables. In: Classification and Related Methods of Data Analysis, H.H. Bock (Ed.), Nort-Holland, 181-186.

Ben-Hur, A., Elisseeff, A., Guyon, I. 2002. A Stability based Method for Discovering Structure in Clustered Data. In Pacific Symposium on Biocomputing, 6-17.

Bertoni, A., Valentini, G. 2007. Model Order Selection for Bio-Molecular Data Clustering.BMC Bioinformatics 2007, 8 (Suppl.3).

Cheng, R., Milligan, G. W. 1996. Measuring the Influence of Individual Data Points in a Cluster Analysis. *Journal of Classification,* 315-35.

Everitt, B. S. 1993. Cluster Analysis (3 ed.. Edward Arnold).

Fridlyand J., Dudoit, S. 2001. Applications of Resampling Methods to Estimate the Number of Clusters and to Improve the Accuracy of a Clustering Method. Technical Report 600, Department of Statistics, University of California, Berkeley.

Goodman, L. A. and Kruskal, W.H. 1954. Measures of Association for Cross-Classifications. *Journal of the American Statistical Association*, 49, 732-64.

Hennig, C. 2004. A General Robustness and Stability Theory for Cluster Analysis. Preprint 2004-07, Fachbereich Mathematik - SPST, Hamburg.

Jain, A. K. and Dubes, R. C. 1988. Algorithms for Clustering Data. NJ Prentice Hall.

Lange, T., Roth, V., Braun, M., Buhmann, 2004. Stability based Validation of Clustering Solutions. *Neural Computation,* 16(6):1299-1323.

Lerman, I. C. 1981. Classification et Analyse Ordinale des Données. Paris, Dunod.

Levine, E., Domany, E. 2001. Resampling Method for Unsupervised Estimation of Cluster Validity. *Neural Computation,* 13(11): 2573-2593.

Mann, H. B., Whitney, D. R. 1947. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 50-60.

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: an evaluation. *Computer Journal*, 20, pp. 359-363.

Mojena, R. and Wishart, D. 1980. Stopping rules for Ward´s clustering method. Proceedings of COMPSTAT 1980. Wurzburg, Germany: Physika-Verlag.

Nicolau, F.C. 1983. Cluster analysis and distribution function, *Methods of Operations Research*, vol. 45, pp. 431-433.

Nicolau, F.C. and Bacelar-Nicolau, H. 1998. Some trends in the classification of variables. In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., &Baba, Y. Eds.), Data Science, Classification, and Related Methods. Springer-Verlag, pp. 89-98.

Rousseeuw, P. J. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computation and Applied Mathematics*, 20, 53-65.

Silva O., Bacelar-Nicolau, H., Nicolau, F. 2009. Como Avaliar a Consistência dos Resultados de uma Análise Classificatória Hierárquica. In: Oliveira, I. et al. Eds. Actas do XVI Congresso Anual da Sociedade Portuguesa de Estatística, 2008, Edições S.P.E., 661-672.

Silva O., Bacelar-Nicolau, H., Nicolau, F.C. 2007. Utilização da Análise Classificatória para Avaliar as Atitudes/Crenças em Relação à Estatística de Alunos da Área de Ciências Sociais e Humanas. In: Ferrão, M. et al. Eds. Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística, 2006, Edições S.P.E, 751-759.

Silva, O. 2011. Contributions to the Evaluation and Comparison of Partition in Cluster Analysis. Ph.D. Thesis (in Portuguese), Universidade dos Açores.

Silva, O., Bacelar-Nicolau, H., Nicolau, F., C. 2012. A global Approach to the Comparison of Clustering Results. Biometrical Letters, 49(2), 135-147.

Smolkin, M., Ghosh, D. 2003. Cluster Stability Scores for Microarray Data in Cancer Studies. *BMC Bioinformatics*, 36 (4).

Sousa, Á., Nicolau, F.C., Bacelar-Nicolau, H., Silva, O. 2014. Cluster Analysis using Affinity Coefficient in order to identify religious beliefs Profiles. *European Scientific Journal (ESJ),* vol. 3 (Special edition), 252 – 261.

****