Adv Data Anal Classif
DOI 10.1007/s11634-009-0057-4

REGULAR ARTICLE

Comparing partitions of two sets of units based on the same variables

Genane Youness · Gilbert Saporta

Received: 13 June 2008 / Revised: 27 March 2009 / Accepted: 27 November 2009

Abstract We propose a procedure based on a latent variable model for the comparison of two partitions of different units described by the same set of variables. The null hypothesis here is that the two partitions come from the same underlying mixture model. We define a method of "projecting" partitions using a supervised classification method: once one partition is taken as a reference; the individuals of the second data set are allocated to the clusters of the reference partition; it gives two partitions of the same units of the second data set: the original and the projected one and we evaluate their difference by usual measures of association. The empirical distributions of the association measures are derived by simulation.

Keywords Rand index · Redundancy index · Discriminant analysis · Latent classes · Partitions

Mathematics Subject Classification (2000) 62H30

1 Introduction

The need of comparing partitions of two different sets of units based on the same questionnaire appears frequently in periodic opinion or market surveys when the questionnaire is asked to different samples.

G. Youness (✉)
Department of Statistics, ISAE-Institut des Sciences Appliquées et Economiques and CEDRIC, CNAM, BP 11-4661, Beirut, Lebanon
e-mail: genane.youness@cnam.fr

G. Saporta
Chaire de Statistique Appliquée and CEDRIC, CNAM, 292 rue Saint Martin, 75003 Paris, France
e-mail: gilbert.saporta@cnam.fr

Published online: 04 January 2010

If all variables are numerical, usual tests for comparing two multivariate distributions might be used such as: Rosenbaum test (2005), nonparametric tests, proportion test of the partition's classes using the chi-square test, Mahalanobis test for comparing the class means of both partitions.

In this paper, a new method of comparing partitions coming from the same variables by "projection" is proposed. Discriminant analysis is applied to one of the partition and the units of the other partition are reclassified. The comparison is done by using association indices under the hypothesis of identical partitions (cf., Youness and Saporta 2004a,b), such as Rand index introduced by Hubert and Arabie (1985), the redundancy index RI introduced by Stewart and Love (1968), and the τ_b index of Goodman and Kruskal (1979).

In order to obtain the distribution of the indices under the hypothesis of identical partitions, we simulate partitions coming from a common latent class model (more precisely a latent profile model (Vermunt and Magidson 2002), since we only deal with p observed numerical variables), which is a particular mixture model. An algorithm has been developed using SAS software in order to check the relevance of our approach. The methodology is then applied to artificial and real data sets. Here, the application deals only with the case of numerical variables.

The paper is organized as follows: Sect. 2 is devoted to define the model used to find the partitions under the null hypothesis: "the partitions are identical". Section 3 describes briefly several indices of agreement. Section 4 gives a detailed description of the procedure of projecting partitions. Finally, we illustrate our methodology on several artificially generated data sets in Sect. 5, and on a real data set in Sect. 6.

2 A latent profile model to define the null hypothesis

A natural idea is to decide that two partitions of the same units do not differ significantly if a measure of agreement is larger than a critical value. Thus, we need to know the probability distribution, even approximated, of an agreement association measure under some null hypothesis. The null hypothesis of independence is inoperative because rejecting the independence does not mean that the partitions are nearly identical. We need to define more precisely the hypothesis H_o where both partitions are identical. As soon as the association measure is larger than some critical value, the hypothesis that the difference between partitions is due to sampling is not rejected at a certain probability level.

We define the hypothesis H_o: "the two partitions are identical", by saying that the units come from the same underlying partition $\mathcal{P}$, and that the two observed partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ are noised realizations of the common partition $\mathcal{P}$. The latent class model provides a nice way of generating data coming from the same underlying partition and has been used by Krieger and Green (1999) in their consensus partition research.

The latent class model corresponds to the following mixture model with local independence:

$$f(x) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} f_k(x_j|k)$$

where π_k $(k = 1, \ldots, K)$ are the class proportions and X is the random vector of observed variables where the components x_j are independent in each class, $f(x)$ is the density of X, and $f_k(x_j|k)$ is the density of x_j in the class k (Bartholomew and Knott 1999).

The latent profile model is used here to generate data and not to estimate parameters (McLachlan 2000, 2004). As usual we generate data sets coming from f as follows. We first generate frequencies n_k from a multinomial distribution with probabilities π_k and then we generate observations in each class according to the local independence model; in other words we choose a normal mixture model with independent components in each class.

3 Some measures of agreement between two partitions of the same units

In previous publications (Youness and Saporta 2004a,b), we have investigated properties of Rand, Mc Nemar, Jaccard, RV-coefficient, JV-index, Cohen's kappa, and Popping's D_2 indices.

We will focus here on the Rand index with its asymmetric version proposed by Chavent et al. (2001) and the redundancy index RI introduced by Stewart and Love (1968).

3.1 Notation

Consider two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ of the same n units with p and q classes, respectively. If $\mathbf{K}_1$ and $\mathbf{K}_2$ are the corresponding $n \times p$ and $n \times q$ disjunctive tables of indicator variables and $\mathbf{N}$ the contingency table with elements $n_{ij}(i = 1, \ldots, p; j = 1, \ldots, q)$, we have:

$$\mathbf{N} = \mathbf{K}_1'\mathbf{K}_2.$$

Each partition $\mathcal{P}_r$ is also characterized by the $n \times n$ paired comparison table $\mathbf{C}^r$ with general term $c^r_{ii'}$:

$$c^r_{ii'} = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ are in the same class of } \mathbf{\Pi}_r \\ 0 & \text{otherwise} \end{cases}$$

we have $\mathbf{C}^1 = \mathbf{K}_1\mathbf{K}_1'$ and $\mathbf{C}^2 = \mathbf{K}_2\mathbf{K}_2'$.

Given n units, $n(n-1)/2$ different pairs of units can be compared. When both partitions assign a pair of units to the same cluster, we have a positive agreement. The number of pairs in positive agreement is denoted a. When both partitions assign a pair of units to different clusters, we have a negative agreement. The number of pairs in negative agreement is denoted b. The number of pairs belonging to the same cluster of the partition $\mathcal{P}_1$ and to different clusters of $\mathcal{P}_2$ is denoted by c. The number of pairs belonging to different clusters of $\mathcal{P}_1$ and to the same clusters of $\mathcal{P}_2$ is denoted by d.

Table 1 Illustration of the four cases

$\mathcal{P}_1 \backslash \mathcal{P}_2$	Same class	Different class
Same class	a	c
	Agreement (same)	Disagreement
Different class	d	b
	Disagreement	Agreement (different)

Table 1 summarizes the four cases.

Let $A = a + b$ be the total numbers of agreements (negative and positive agreements).

3.2 Rand index

Symmetric Rand index

The Rand index, similar to Kendall measure (1938), is the proportion of agreements (positive and negative):

$$R = \frac{2A}{n(n-1)}$$

It may be proved that:

$$A = \binom{n}{2} + \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}^2 - \frac{1}{2}\left[\sum_{i=1}^{p} n_{i.}^2 + \sum_{j=1}^{q} n_{.j}^2\right]$$

where n_{ij}, denoting the frequency of the pair (i, j), is the generic term of the cross tabulation of $\mathcal{P}_1$ and $\mathcal{P}_2$.

We have $0 \leq R \leq 1$ and $R = 1$ if $\mathcal{P}_1 = \mathcal{P}_2$.

To take into account all n^2 pairs including the identical ones, the Marcotorchino modified version of the Rand index, is used (cf. Marcotorchino and El Ayoubi 1991). This modified version of Rand index is expressed as:

$$\text{Rand} = \frac{2\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}^2 - \sum_{i=1}^{p} n_{i.}^2 - \sum_{j=1}^{q} n_{.j}^2 + n^2}{n^2}$$

We do not use here the correction for chance of the Rand index, introduced by Hubert and Arabie (1985), since it gives negative values when both partitions are not close enough.

Asymmetric version of Rand index

If $p > q$, $\mathcal{P}_1$ is a refinement of $\mathcal{P}_2$ when two elements in the same cluster of $\mathcal{P}_1$ are also in the same cluster of $\mathcal{P}_2$. Then we can measure the degree of inclusion of partition $\mathcal{P}_1$ in $\mathcal{P}_2$ by the asymmetric version of Rand index, proposed by Chavent et al. (2001).

For all n^2 pairs, the asymmetric version of Rand index is defined by:

$$R_A(\mathcal{P}_1, \mathcal{P}_2) = 1 + \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\binom{n_{ij}}{2} - \sum_{i=1}^{p}\binom{n_{i.}}{2}}{\binom{n}{2}}$$

R_A takes values in [0, 1] and $R_A = 1$ if $\mathcal{P}_2 \subseteq \mathcal{P}_1$.

For this study, we use the analogous version with all n^2 pairs (Youness and Saporta 2004a):

$$R_A(\mathcal{P}_1, \mathcal{P}_2) \quad = \frac{n^2 + \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}^2 - \sum_{i=1}^{p} n_{i.}^2}{n^2}$$

It should be noted that if the two partitions have the same number of clusters, the asymmetric version of Rand index is not equal to the initial Rand index: the difference is due to the fact that the asymmetric Rand index is a prediction measure depending on the choice of a reference partition.

3.3 Redundancy index and asymmetric τ_b

Proposed by Stewart and Love (1968) RI is an asymmetric index defined as:

$$\mathrm{RI}(\mathbf{X}_1, \mathbf{X}_2) = \frac{\text{trace } (\mathbf{W}_{12}\mathbf{W}'_{22}\mathbf{W}_{21})}{\text{trace } (\mathbf{W}_{11})}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are two numerical data tables of the same units and $\mathbf{W}_{ij} = \mathbf{X}_i\mathbf{X}'_j$.

RI is a weighted average of the squared multiple correlation coefficients between all pairs of variables of $\mathbf{X}_1$ and $\mathbf{X}_2$. It is a measure of the quality of prediction of $\mathbf{X}_1$ by $\mathbf{X}_2$ and represents the proportion of the explained variance in the regression of $\mathbf{X}_1$ by $\mathbf{X}_2$. $0 \leq \mathrm{RI} \leq 1$ and RI is equal to the square multiple correlation coefficient when $p = 1$ and $q > 1$.

In the case of numerical variables, Lazraq and Cleroux (2002) have proposed to test the null hypothesis that RI is equal to zero and derived a method for selecting predictors in multivariate regression.

When $\mathbf{X}_1$ and $\mathbf{X}_2$ are the tables of indicator variables of two partitions, RI is equal to the asymmetric dependence index $\tau_b$of Goodman and Kruskal (1979), (see Saporta 2006).τ_b measures the proportional reduction in error (PRE) of a prediction rule based on conditional probabilities of assigning an unit to a cluster, compared to the prediction based on marginal probabilities. For two partitions, $\mathcal{P}_1$ and $\mathcal{P}_2$, τ_b is:

$$\tau_b = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}^2}{n.n_{i.}} - \sum_{j=1}^{q}\left(\frac{n_{.j}}{n}\right)^2}{1 - \sum_{j=1}^{q}\left(\frac{n_{.j}}{n}\right)^2}$$

$\tau_b = 0$ in case of independence and $\tau_b = 1$ in case of perfect linear relationship.

We will use this asymmetric index to compare partitions with different number of clusters, coming from the same data set. A high value of this index may lead to the conclusion that partitions are almost identical.

4 Projecting partitions

In order to compare two partitions of different units described by the same set of variables, we use a "projection" technique which basically consists to boil down to the case of comparing two partitions of the same units which has been already dealt with in many publications (eg. Hubert and Arabie 1985; Krieger and Green 1999; Overall and Magee 1992; Saporta and Youness 2002).

It consists in allocating the units of the second set to the clusters defined by the reference partition using some discriminant analysis technique. Thus, for each unit of the second data set I_2 we have both the class in its natural partition (found by e.g. a classical *k*-means algorithm) and the predicted class (found by discriminant analysis), in the reference partition. So the problem will come down to compare two partitions of the same units of I_2: $\mathcal{P}_2$ and $\mathcal{P}'_2$. $\mathcal{P}_2$ comes out from a direct cluster analysis of I_2; $\mathcal{P}'_2$ comes out according to classes of $\mathcal{P}_1$ defined on I_1 using a supervised classification method (cf. Fig. 1).

Many classification methods may be used for such a purpose (linear, quadratic, neural nets, decision trees etc., see McLachlan 2004). For the sake of simplicity, a linear multigroup discriminant analysis which is known to be optimal for normal distribution with equal covariance matrices (Hand 1981) will be used in this paper.

The projection of a partition on the other one is performed as follows:

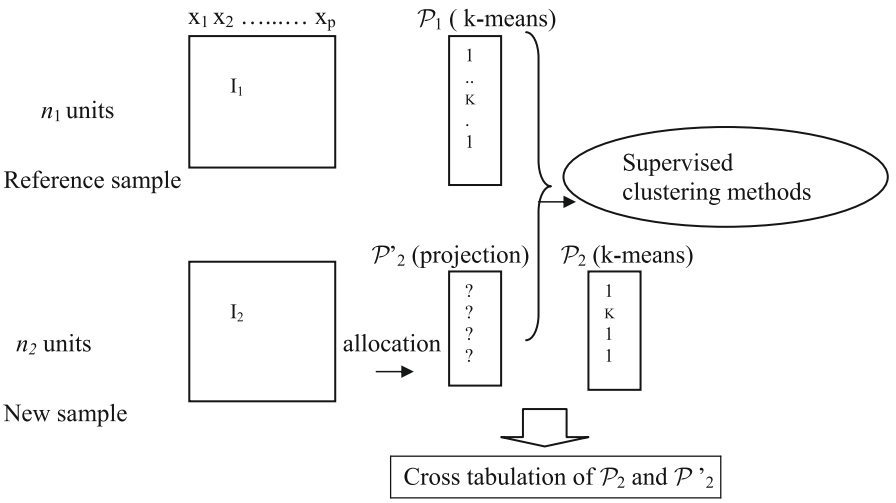

Fig. 1 Scheme of projecting a partition on another one

5 Simulating the projection method and the sampling distributions of agreement indices

In order to study the behaviour of the agreement indices under the null hypothesis of a common latent partition we use the following algorithm:

1. Generate cluster sizes $n_1, n_2, \ldots, n_K$ from a multinomial distribution M($n, \pi_1, \pi_2, \ldots, \pi_K$). For each cluster i, generate n_i values from a random normal vector with p independent components. So the first data set I_1 of n_1 units is obtained by simulation of a normal mixture model with p variables. For this data an associated partition $\mathcal{P}_1$ found by k-means with k_1 classes is chosen as a reference ($k_1 = 1, \ldots, K$, is the number of clusters of $\mathcal{P}_1$).
2. The same independent normal variables are used to generate the second data set I_2 of n_2 units. The second set I_2 of n_2 units is obtained according to the same latent profile model with an associated partition $\mathcal{P}_2$ found by k-means with k_2 classes ($k_2 = 1, \ldots, K$, it is the number of clusters of $\mathcal{P}_2$). The data base I with $n_1 + n_2 = n$ units is obtained by merging I_1 and I_2.
3. The units of the second set I_2 are allocated to the k_1 classes of $\mathcal{P}_1$ by applying a linear discriminant analysis on I_2 to obtain a new partition $\mathcal{P}_2'$. To find the projected partition $\mathcal{P}_2'$, the procedure DISCRIM of SAS (9.1) has been used: a multiclass linear discriminant analysis is performed on $\mathcal{P}_1$. Then each unit of I_2 is classified into one of the class of the reference partition $\mathcal{P}_1$.
4. Compute association indices to measure the difference between the partitions $\mathcal{P}_2$ and $\mathcal{P}_2'$ of the same set I_2. When partition $\mathcal{P}_2$ has a number of clusters k_2 different from the number of clusters k_1 of $\mathcal{P}_1$, we use an asymmetric version of Rand index or of the redundancy index.

Now in order to get the sampling distributions of the indices we split I at random into two sets of sizes n_1 and n_2 and repeat steps 3 and 4 for a number of times.

6 Some simulations

We simulate 500 samples of 1,000 units coming from a latent profile model with 4 classes and 4 variables which are normally distributed in each class.

The parameters of the normal distributions are chosen such that for every variable $j = 1, \ldots, 4$, the absolute value of the difference between the means $m_{kj}, m_{k'j}$ of the normal distributions of two different classes $k, k' = 1, \ldots, 4$, $k \neq k'$, is larger than its standard deviation by one and a half:

$$\left|m_{kj} - m_{k'j}\right| > 1.5\sigma_j \quad \text{for } j = 1, 2, 3, 4 \quad \text{and} \quad k, k' = 1, 2, 3, 4, \ k \neq k'.$$

6.1 First example

The first choice of the parameters for the independent normal variables is given in Table 2.

Table 2 First simulated example: parameters of the normal mixture model

Class 1	Class 2	Class 3	Class 4
X1 $N(1.2, 1.5)$	X1 $N(-2, 1.5)$	X1 $N(5, 1.5)$	X1 $N(8, 1.5)$
X2 $N(-10, 2.5)$	X2 $N(0, 2.5)$	X2 $N(-417, 2.5)$	X2 $N(3.8, 2.5)$
X3 $N(6, 3.5)$	X3 $N(12, 3.5)$	X3 $N(13, 3.5)$	X3 $N(-5, 3.5)$
X4 $N(-20, 4.5)$	X4 $N(-12, 4.5)$	X4 $N(0, 4.5)$	X4 $N(7, 4.5)$

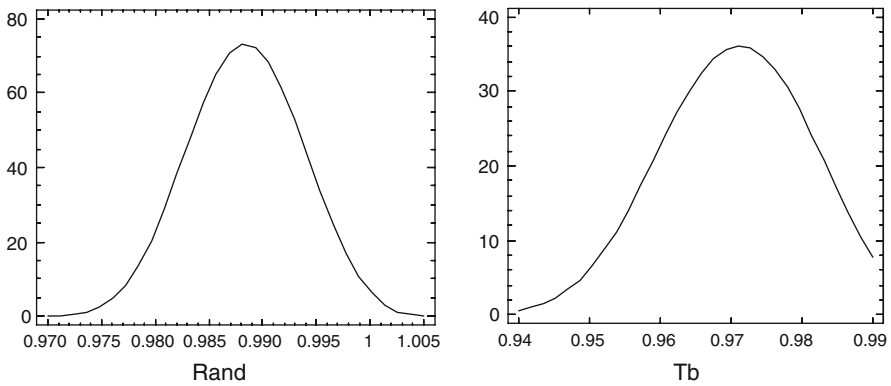

Fig. 2 Kernel density estimates of Rand and τ_b for 500 iterations

Table 3 Descriptive statistics for Rand and τ_b for the first choice of parameters

	Rand	τ_b
Frequency	500	500
Mean	0.988367	0.970632
Mode	0.992	0.974
Variance	0.0000132944	0.0000823894
Standard deviation	0.00364615	0.00907686
Minimum	0.977	0.943
Maximum	0.996	0.989
Range	0.019	0.046
Skewness	−1.09906	−1.64516
Kurtosis	−1.81607	−1.95615

The algorithm described in Sect. 5 is repeated 500 times which enables us to get approximate distributions of τ_b and of the Marcotorchino modified version of the Rand index (Fig. 2).

The values of the Rand index vary from 0.977 and 0.996. The most frequent value is 0.992 and its mean is equal to 0.988. Goodman–Kruskal index τ_b takes its values between 0.943 and 0.989 with a mode equal to 0.974. The distribution has a mean equal to 0.97 (Table 3).

Table 4 gives the 95% confidence intervals of the mean for these indices.

Table 4 Confidence intervals for the Rand index and τ_b for the first simulated example

	Mean	Standard error	Lower limit	Upper limit
Rand	0.988367	0.000163061	0.988047	0.988687
τ_b	0.970632	0.000405929	0.969834	0.97143

Table 5 Second simulated example: parameters of the normal mixture model

Class 1	Class 2	Class 3	Class 4	Class 5
X1 $N(1.2, 1.5)$	X1 $N(-2, 1.5)$	X1 $N(-5, 1.5)$	X1 $N(-8, 1.5)$	X1 $N(8, 1.5)$
X2 $N(4, 2.5)$	X2 $N(-4, 2.5)$	X2 $N(-10, 2.5)$	X2 $N(-15, 2.5)$	X2 $N(8.2, 2.5)$
X3 $N(7, 3.5)$	X3 $N(-6, 3.5)$	X3 $N(-13, 3.5)$	X3 $N(-20, 3.5)$	X3 $N(20, 3.5)$
X4 $N(10, 4.5)$	X4 $N(-10, 4.5)$	X4 $N(-20, 4.5)$	X4 $N(-30, 4.5)$	X4 $N(30, 4.5)$

Under the null hypothesis of identical partitions, these indices have values around 0.988 thus the decision that the two partitions do not differ significantly could be taken if the index is larger than a critical value. Here the estimated critical values at 5% are 0.954 for RI and 0.982, for Rand.

6.2 Second example

A second choice is studied in Table 5 using the normal mixture model:

In this section, we consider the case where different number of clusters are used for the two partitions: $\mathcal{P}_1$ with 5 clusters while $\mathcal{P}_2$ has 2 clusters. As before partition $\mathcal{P}_2'$ is obtained by projection on $\mathcal{P}_1$.

The association indices τ_b and the asymmetric version of Rand index for both partitions are calculated for 500 repetitions.

The redundancy index RI takes its values between 0.80 and 0.99 with a mode equal to 0.97. The distribution has a mean equal to 0.868. The values of the asymmetric version of Rand index vary from 0.95 to 0.998. So under the null hypothesis of identical partitions, the 5% critical value of the redundancy index RI is equal to 0.81 and the corresponding critical value of the asymmetric version of Rand is equal to 0.97 (Fig. 3).

Unfortunately no universal critical values for distribution of agreement indices can be obtained: they depend on the number of clusters, of observations and of cluster separation. In our algorithm, this problem is solved by finding the critical value which corresponds to the distribution of each index found under the hypothesis of identical partitions.

In the next section, a generic algorithm is implemented and applied on a real data set, showing how our method can be generally used.

7 Real data example

We investigated the performance of our technique on the following real data set: a survey about conditions of life and expectations of a French sample (Lebart 1987).

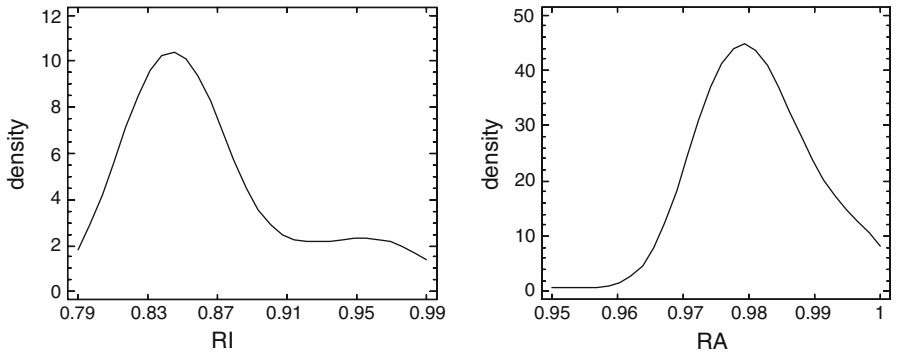

Fig. 3 Kernel density estimates of the redundancy index and the asymmetric Rand index with different number of clusters

The data set contains 1,000 individuals and 52 variables. We selected 14 variables about quality of life, opinions about marriage, family and children. After eliminating missing values, we have 624 individuals: 315 men and 309 women.

The goal is to compare men's and women's partitions. Therefore an empirical sampling distribution of association index has to be found under the hypothesis of "identical partitions". We obtain the distribution of the redundancy index, and the Rand index and its asymmetric version, after 500 iterations using the following steps:

- Do random sample of the data set I with the same variables and split it at random into two data sets I_1 and I_2 of 312 units each.
- Find the partition $\mathcal{P}_1$ of the data set I_1 with p clusters, by k-means.
- The units of the second set I_2 are allocated to the p clusters of $\mathcal{P}_1$ by applying a linear discriminant analysis on I_2 to obtain a new partition $\mathcal{P}_2'$, using the procedure DISCRIM of SAS.
- Find another partition $\mathcal{P}_2$ of the second data base I_2 with$q$ clusters by k-means.
- Cross classify the two partitions $\mathcal{P}_2$ and $\mathcal{P}_2'$, and compute association measures.

For this data set, we chose for both partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ the same number of clusters $p = q = 4$ because they give the best partitions of the trees owing to a maximization of the inter-class inertia.

The distribution of the indices under the null hypothesis of identical partitions is approximated after 500 repetitions of random splits (Table 6, Fig. 4).

The 95 % confidence intervals of the mean, for these indices are in Table 7.

The lower 5% quantile under the hypothesis of identical partitions is equal to 0.650 for Rand, 0.85 for R_A and 0.35 forτ_b.

To compare the partition of men and women, we run the algorithm after sorting the variable sex and dividing the real data set according to this variable.

We projected men's partition on women's one and vice versa which gives two values forτ_b: 0.185, 0.2582 and two values for the Rand index: 0.6134, 0.6466. Since they are lower than the corresponding critical values, we may conclude that both partitions are different.

Table 6 Descriptive statistics for the Rand, R_A and τ_b indices for the real data set

	Rand	R_A	τ_b
Frequency	500	500	500
Mean	0.721302	0.850708	0.35129
Mode	0.71	0.86	0.3
Variance	0.0025087	0.001526	0.011745
Standard deviation	0.050087	0.039064	0.108374
Minimum	0.592	0.685	0.1
Maximum	0.88	0.933	0.7
Range	0.288	0.248	0.6
skewness	6.21384	-4.89316	6.18305
Kurtosis	1.64771	3.88624	0.931154

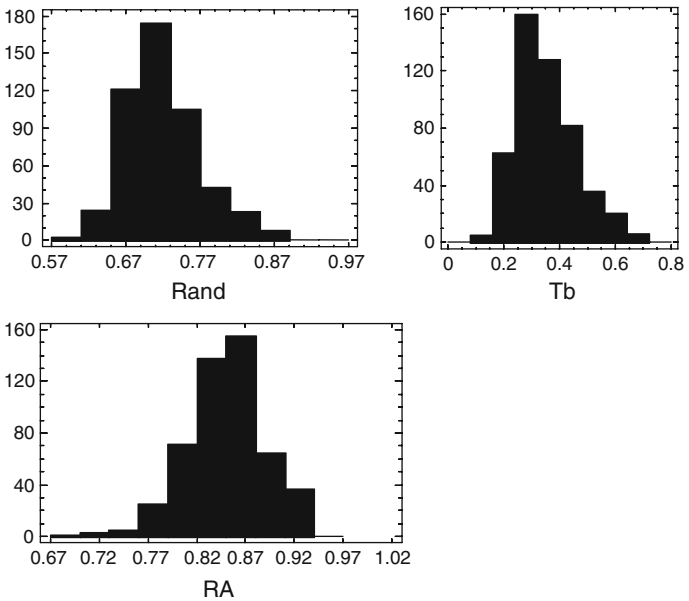

Fig. 4 Distributions of the Rand, R_A and τ_b indices in 500 iterations

Table 7 Confidence intervals of the Rand, R_A and τ_b indices for the real data set

	Mean	Standard error	Lower limit	Upper limit
Rand	0.721302	0.00223996	0.716901	0.725703
R_A	0.850708	0.001747	0.847276	0.85414
τ_b	0.35129	0.00484665	0.341768	0.360812

8 Conclusion

We have presented a new method of comparing partitions coming from two sets of objects with the same variables based on a projection of partitions, through discriminant analysis.A latent profile model has been used to solve the problem of comparing partitions by simulation. The comparison of partitions has been done using agreement measures such as the Rand index and the redundancy index. The redundancy index RI that allows for testing similarity between two partitions with different number of clusters, has been studied.

We have proposed a simulation procedure to obtain critical values for each index. They are data dependent on the number of clusters, their proportions and their separation.

Our methodology has been illustrated on a real data set. The results obtained in the simulation study and in the empirical analysis are encouraging, showing the usefulness of the proposals and the efficiency of the algorithm.

However, further studies are needed concerning for instance the meaning of classes in terms of the variables (external and internal information).

References

Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis. Arnold, London

Chavent M, Lacomblez C, Patouille B (2001) Critère de Rand asymétrique. In: Proceedings SFC 2001, 8eme Rencontres de la Société Francophone de la Classification, Pointe à Pitre, France, pp 82–88.

Goodman L, Kruskal W (1979) Measures of association for cross-classifications. Springer, New York

Hand DJ (1981) Discrimination and classification. Wiley, London

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–198

Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

Krieger A, Green PA (1999) Generalized Rand-index method for consensus clustering of separate partitions of the same data base. J Classif 16:63–89

Lazraq A, Cleroux R (2002) Inférence robuste sur un indice de redondance. Revue de Statistique Appliquée 4:39–54

Lebart L (1987) Conditions de vie et aspirations des français, évolution et structure des opinions de 1978 à 1984. Futuribles 1:25–26

McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York

McLachlan GJ (2004) Discriminant analysis and statistical pattern recognition. Wiley, New York

Marcotorchino JF, El Ayoubi N (1991) Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. Revue de Statistique Appliquée 2:25–46

Overall JE, Magee K (1992) Estimating individual rater reliabilities. Appl Psychol Meas 16(1):77–85

Rosenbaum PR (2005) An exact distribution-free test comparing two multivariate distributions based on adjacency. J Royal Stat Soc B 67(4):515–530

Saporta G (2006) Probabilités analyse des données et statistique, 2eme edn. Technip, Paris

Saporta G, Youness G (2002) Comparing two partitions: some proposals and experiments. In: Wolfgang H (ed) Proceedings in computational statistics, Physica-Verlag, Berlin, Germany, pp 243–248

Stewart D, Love W (1968) A general canonical correlation index. Psychol Bull 70:160–163

Vermunt JK, Magidson J (2002) Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL (eds) Applied latent class analysis. Cambridge University Press, Cambridge, pp 89–106

Youness G, Saporta G (2004) Une méthodologie pour la comparaison de partitions. Revue de Statistique Appliquée LII(1):97–120

Youness G, Saporta G (2004) Some measures of agreement between close partitions. J Student 5(1):1–12