ISCTE IUL
Instituto Universitário de Lisboa

Escola de Gestão
Departamento de Métodos Quantitativos para Gestão e Economia

On Clustering Stability

Maria José de Pina da Cruz Amorim

Compilação de artigos especialmente elaborada para obtenção do grau de

Doutor em Métodos Quantitativos

Orientadora:
Doutora Maria Margarida Guerreiro Martins dos Santos Cardoso, Professora Associada com Agregação, ISCTE – Instituto Universitário de Lisboa

Dezembro, 2015

ISCTE IUL
Instituto Universitário de Lisboa

Escola de Gestão
Departamento de Métodos Quantitativos para Gestão e Economia

On Clustering Stability

Maria José de Pina da Cruz Amorim

Compilação de artigos especialmente elaborada para obtenção do grau de

Doutor em Métodos Quantitativos

Júri:
Doutora Elizabeth de Azevedo Reis, Professora Catedrática, ISCTE- Instituto Universitário de Lisboa
Doutor Mário Alexandre Teles de Figueiredo, Professor Catedrático, Instituto Superior Técnico de Lisboa
Doutora Teresa Paula Costa Azinheira Oliveira, Professora Auxiliar, Universidade Aberta
Doutora Ana Alexandra Antunes Figueiredo Martins, Professora Adjunta, Instituto Superior de Engenharia de Lisboa
Doutora Paula Alexandra Barbosa da Conceição Vicente Duarte, Professora auxiliar com Agregação, ISCTE – Instituto Universitário de Lisboa
Doutora Maria Margarida Guerreiro Martins dos Santos Cardoso, Professora Associada com Agregação, ISCTE – Instituto Universitário de Lisboa

Dezembro, 2015

À minha mãe e ao João

AGRADECIMENTOS

Agradeço a todas as pessoas que directa ou indirectamente me apoiaram e incentivaram ao longo deste trabalho de investigação.

Agradeço especialmente à Professora Margarida Cardoso por tudo aquilo que me ensinou, pelo seu empenho, dedicação e pela sua orientação, sem a qual este trabalho não teria sido possível.

Agradeço ao Instituto Superior de Engenharia de Lisboa o facto de ter pago parte das propinas do doutoramento e me ter concedido equiparação a bolseiro em tempo integral durante um semestre.

Abstract

This work is dedicated to the evaluation of the stability of clustering solutions, namely the stability of crisp clusterings or partitions. We specifically refer to stability as the concordance of clusterings across several samples. In order to evaluate stability, we use a weighted cross-validation procedure, the result of which is summarized by simple and paired agreement indices values. To exclude the amount of agreement by chance of these values, we propose a new method – IADJUST – that resorts to simulated cross-classification tables. This contribution makes viable the correction of any index of agreement.

Experiments on stability rely on 540 simulated data sets, design factors being the number of clusters, their balance and overlap. Six real data with a priori known clusters are also considered. The experiments conducted enable to illustrate the precision and pertinence of the IADJUST procedure and allow to know the distribution of indices under the hypothesis of agreement by chance. Therefore, we recommend the use of adjusted indices to be common practice when addressing stability. We then compare the stability of two clustering algorithms and conclude that Expectation-Maximization (EM) results are more stable when referring to unbalanced data sets than K means results. Finally, we explore the relationship between stability and external validity of a clustering solution. When all experimental scenarios' results are considered there is a strong correlation between stability and external validity. However, within a specific experimental scenario (when a practical clustering task is considered), we find no relationship between stability and agreement with ground truth.

Keywords: Adjusted índices of agreement, Clustering evaluation, External evaluation, Clustering stability.

JEL Classification: C100; C150; C380

Resumo

Este trabalho é dedicado à avaliação da estabilidade de agrupamentos, nomeadamente de partições. Consideramos a estabilidade como sendo a concordância dos agrupamentos obtidos sobre diversas amostras. Para avaliar a estabilidade, usamos um procedimento de validação cruzada ponderada, cujo resultado é resumido pelos valores de índices de concordância simples e pareados. Para excluir, destes valores, a parcela de concordância por acaso, propomos um novo método - IADJUST - que recorre à simulação de tabelas cruzadas de classificação. Essa contribuição torna viável a correção de qualquer índice de concordância.

A análise experimental da estabilidade baseia-se em 540 conjuntos de dados simulados, controlando os números de grupos, dimensões relativas e graus de sobreposição dos grupos. Também consideramos seis conjuntos de dados reais com classes a priori conhecidas. As experiências realizadas permitem ilustrar a precisão e pertinência do procedimento IADJUST e conhecer a distribuição dos índices sob a hipótese de concordância por acaso. Assim sendo, recomendamos a utilização de índices ajustados como prática comum ao abordar a estabilidade. Comparamos, então, a estabilidade de dois algoritmos de agrupamento e concluímos que as soluções do algoritmo Expectation Maximization são mais estáveis que as do K-médias em conjuntos de dados não balanceados. Finalmente, estudamos a relação entre a estabilidade e validade externa de um agrupamento. Agregando os resultados dos cenários experimentais obtemos uma forte correlação entre estabilidade e validade externa. No entanto, num cenário experimental particular (para uma tarefa prática de agrupamento), não encontramos relação entre estabilidade e a concordância com a verdadeira estrutura dos dados.

Palavras-chave: índices de concordância ajustados, avaliação de agrupamentos, validação externa de agrupamentos, estabilidade

Sistema de Classificação JEL: C100; C150; C380

Sumário Executivo

Este trabalho é dedicado à avaliação da estabilidade de agrupamentos, nomeadamente de partições. Consideramos a estabilidade como sendo a concordância dos agrupamentos obtidos sobre diversas amostras. Para avaliar a estabilidade, usamos um procedimento de validação cruzada ponderada, cujo resultado é resumido pelos valores de índices de concordância simples e pareados. Para excluir, destes valores, a parcela de concordância por acaso, propomos um novo método – IADJUST. Segundo este método são simuladas tabelas de classificação cruzada, obedecendo estas aos totais marginais da tabela de classificação cruzada observada, ou seja a tabela referida às partições que se pretendem comparar. A geração das tabelas de classificação cruzada é feita de acordo com o modelo Hipergeométrico. Utiliza-se a linguagem de programação R para implementar o método. A contribuição do IADJUST torna viável a correção de qualquer índice de concordância, superando limitações de métodos analíticos e aproximativos conhecidos à data.

A análise experimental da estabilidade baseia-se em 540 conjuntos de dados simulados, controlando os números de grupos (2, 3 e 4), dimensões relativas (grupos balanceados e não balanceados) e graus de sobreposição dos grupos (grupos bem, moderadamente e fracamente separados). As bases de dados simulados são geradas de acordo com modelos de mistura finita de normais multivariadas, utilizando o package MixSim para o efeito. A análise é também efetuada sobre seis conjuntos de dados reais, com classes a priori conhecidas (disponíveis no UCI Machine Learning Repository)

As experiências realizadas permitem ilustrar a precisão e pertinência do procedimento IADJUST e conhecer a distribuição dos índices sob a hipótese de concordância por acaso (H0). Assim sendo, recomendamos a utilização de índices ajustados como prática comum ao abordar a estabilidade de agrupamentos e, em certos casos, a utilização da mediana em vez da média (sob H0) para este ajustamento.

Comparando a estabilidade de dois algoritmos de agrupamento - K-Médias e Expectation Maximization (EM) – incorporamos o procedimento IDJUST. Verificamos que, nos cenários com grupos de dimensão balanceada as partições obtidas com o K médias são mais estáveis e nos cenários em que os grupos têm dimensões desproporcionadas são mais estáveis as soluções obtidas com o EM.

Finalmente, este trabalho contribui com uma nova perspetiva para uma melhor compreensão da relação entre estabilidade de um agrupamento e a sua validade externa. Agregando os resultados dos cenários experimentais obtemos uma forte correlação entre estabilidade e validade externa. No entanto, num cenário experimental particular (para uma tarefa prática de agrupamento), não encontramos relação entre estabilidade de um agrupamento e a sua concordância com as verdadeira classes. Assim, e embora um agrupamento instável continue a ser, por isso mesmo, indesejável (senão que resultados deveria um analista escolher?), constata-se que não há uma relação credível entre a estabilidade de uma partição e a sua concordância com a verdadeira estrutura dos dados.

Palavras-chave: índices de concordância ajustados, avaliação de agrupamentos, validação externa de agrupamentos, estabilidade

Sistema de Classificação JEL: C100; C150; C380

CONTENTS

CHAPTER 1: INTRODUCTION

On Clustering Validation and Research Objectives

Cluster analysis is the partitioning of a data set into groups (clusters), so that data points within a group are similar to each other and elements of different clusters are dissimilar. A clustering is a set of clusters that covers a data set, commonly a partition.

Cluster analysis can be used in a wide variety of fields such as biology Wu (2011), ecology (Ravera, 2001), medical sciences (Marateb et al., 2014), image and market segmentation (Müller and Hamm, 2014), music (instrument recognition, find musical structure) (Krey et al., 2014), engineering (Todeschini et al., 2012; Krey et al., 2015), and many more. "Solving a clustering problem may also help solving other related problems, such as pattern classification and rule extraction from data" (Vendramin et al., 2010: 209).

"Clustering methods range from those that are largely heuristic to more formal procedures based on statistical models." (Fraley and Raftery, 1998: 578). In any case, when we conduct a cluster analysis the assessment of the quality/validity of the solution obtained is an extremely important issue since:

- Cluster analysis is an unsupervised learning method and so, in practical applications, "there is no ground truth against which we could "test" our clustering results" (Luxburg, 2009): 236);
- Clustering methods tend to generate clusterings even for fairly homogeneous data sets" (Hennig, 2007): 258);
- The clustering solution depends on the method used, different methods emphasizing different aspects of the data and thus enforcing different structure on data;
- Clustering methods‟ parametrization influences the corresponding results since the methods "attempt to define the best partitioning of data set for the given parameters … and not necessarily the „best‟ one that fits the data set." (Halkidi and Vazirgiannis, 2001: 111).

When evaluating a clustering solution two main perspectives can be considered: internal and external.

The **external evaluation** of clustering results resorts to data sets with known cluster structure and measures agreement between this structure and the results obtained. The procedure allows evaluating the ability of algorithms to recover the classes known *a priori*.

In general, external evaluation is carried out both on synthetic data and real data. The first may be resulting from generation of finite mixture models, for example - (Maitra and Melnykov 2010). Real data (with known classes) can be obtained, for example, in the Machine Learning UCI Repository - (Bache and Lichman, 2013).

The **internal evaluation** of clustering results relies in the clusters cohesion and separation. These are inherent properties to the very idea of clustering, reflecting the internal homogeneity of elements in the same group and the isolation of a group when compared with others (heterogeneity between groups). The evaluation of these properties resorts to indices of cohesion-separation, which commonly are ratios between a measure of intra-groups vs. between-groups variation, e.g. (Cardoso et al., 2009).

In the context of internal evaluation we can address clustering stability. Stability has been recognized as a desirable property of a clustering solution – e.g. (Jain and Dubes, 1988; Mirkin, 1996; Gordon, 1999; Everit et al., 2001; Lange et al., 2004; Alizadeh et al., 2014). A clustering solution is said to be stable if it remains fairly unchanged when the clustering process is subject to minor modifications such as alternative parameterizations of the algorithm used, introducing noise in the data or using different samples.

Ben-David and Von Luxburg (2008) warn of a possible misuse of stability noting that the goodness of this property in the evaluation of clustering results is not theoretically well founded: "While it is a reasonable requirement that an algorithm should demonstrate stability in general, it is not obvious that, among several stable algorithms, the one which is most stable leads to the best performance" (p. 379).

This work is dedicated to the evaluation of the stability of clustering solutions, namely the stability of crisp clustering solutions or partitions. We aim at contributing to provide adequate measures of agreement between partitions which should exclude agreement by chance. In addition, the relationship between stability and external validity of a clustering solution is explored in order to improve the understanding of the role of stability when evaluating a clustering solution.

Stability and Cross-Validation

In this work we specifically consider the stability concept which refers to variability/consistency of clustering solutions across several samples. There are several contributions in this domain. Table 1 summarizes some recent works including the specification of **subsampling scheme**, **cross-validation** method and **measures of agreement** quantifying the degree of a clustering solution stability.
In order to obtain the different samples to test clustering stability we can, successively, draw, with replacement or not, a random sample (with size m) of the data set (with size n).

When using bootstrap (the points of each sample drawn with replacement) we usually have $m = n$ and the clustering obtained with the original data set is compared with the clusterings derived from the boostrap samples. The comparison is then based on the points of the data set that are also in the samples. A disadvantage of bootstrap, in the context of clustering, is the possible occurrence of multiple points, which the algorithm can consider as mini-clusters (Monti et al., 2003).

When the points of each sample are drawn without replacement (subsetting) we can decide to obtain overlapping or disjoint samples (splits of the data set). In both cases we must decide about the sample size ($m < n$). It is worthwhile to note that if the training sample size is too small, the structure of the data set can be overlooked. In general, we can consider (as a rule of thumb), a fraction of points of the data set sampled over 0.5 to anticipate the same general clustering structure in the subsamples (Ben-Hur et al., 2002).

When comparing two clustering solutions arising from overlapping samples, the data points which belong to both samples can be used to evaluate the agreement between the clusterings.

Table 1- Some works on stability

References	Context / (Algorithms)	Subsampling scheme	Cross-validation/ similarity on the points common	Measure of agreement
(Cheng and Milligan, 1996)	Measuring the impact of individual data points in a cluster analysis (different algorithms)	The clustering methods were applied to all data sets with a single element at a time left out	Similarity on the points common	Adjusted Rand Gamma Point-biserial statistics
(Ben-Hur et al., 2002)	Estimate the true number of clusters (average-link hierarchical clustering)	Repeated subsampling with a fraction of the data	Similarity on the points common	Fowlkes and Mallows
(Roth et al., 2002)	Find right number of clusters (annealing version of the *k*-means)	Repeated Half-Splits of the data set	Cross-validation	Average of the fraction of differently labeled points of two solutions normalized by the stability of a random predictor
(Dudoit and Fridlyand, 2002)	Find right number of clusters (PAM algorithm)	Repeated splits of the data set into two disjoint subsets	Cross-validation	Fowlkes and Mallows
(Lange et al., 2004)	Find the number of clusters (annealing version of the *k*-means and a path-based clustering)	Repeated splits of the data set into two disjoint subsets	Cross-validation	Normalized Hamming distance
(Hennig, 2007)	A cluster-wise measure of cluster stability (different algorithm)	Non-parametric Bootstraping		Jaccard
(Dolnicar and Leisch, 2010)	Find the final clusters (k-means, neural gas algorithm, binomial mixture model)	Bootstraping		Adjusted Rand
(Pascual et al., 2010)	Find right number of clusters (K-means, Gaussian mixture model, H-Density)	Repeated splits of the data set into two disjoint subsets	Cross-validation	Proposed a cluster stability index based in Mutual Information

Table 1 – Some works on stability (cont)

References	Context / (Algorithms)	Subsampling scheme	Cross-validation/ similarity on the points common	Measure of agreement
(Wang, 2010)	Find right number of clusters (K-means and spectral clustering)	Split, r times, the data into two training sets and one validation set.	Modified cross-validation	Hamming distance
(Fang and Wang, 2012)	Find right number of clusters (K-means, spectral clustering algorithm)	Non-parametric Bootstraping		Hamming distance
(Hennig and Liao, 2013)	Explore the use of formal clustering methods for socio-economic stratification based on mixed-type data with continuous, ordinal and nominal variables	The clustering methods were applied to all data sets with one variable at a time left out	Similarity on the points common	Adjusted Rand
(Alizadeh et al., 2014)	Build a Clustering ensemble	Resampled data sets	similarity on the points common	Normalized Mutual Information
(Krey et al., 2014)	Find right number of clusters (K-means)	Bootstraping		Adjusted Rand
(Krey et al., 2015)	Find right number of clusters (stability with K-means)	Power line (edge of the graph) is cut one after the other to simulate different failure situations The clustering results in those scenarios are compared with the full network graph	similarity on the points common	Adjusted Rand

When evaluating the agreement between two clustering solutions arising from different (disjoint) samples we have to transfer the solutions to a common sample (typically one of the original samples) so that we can measure their agreement. Such transfer may be accomplished by a cross-validation procedure.

The use of cross-validation in the context of clustering evaluation is first proposed by (McIntyre and Blashfield 1980) and (Breckenridge 1989). They import a methodology commonly used in supervised analysis to the domain of unsupervised analysis - Table 2.- At the end of the cross-validation procedure, the value of an index of agreement (obtained in an holdout sample or test sample) between clustering results and supervised classification (by clusters in the training sample), is used as an indicator stability – see Indices of Agreement between partitions, p. 7.

Table 2. - General Cross-Validation procedure

Step	Action	Output
1	Perform training-test sample split	Training and test samples
2	Cluster training sample	Clusters in the training sample
3	Build a classifier using the training sample supervised by clusters' labels; use the classifier in the test sample.	Classes in the test sample
4	Cluster the test sample	Clusters in the test sample
5	Obtain a contingency table between clusters and classes in the test sample and calculate indices.	Indices of agreement values, indicators of stability

In this work, we resort to the weighted cross-validation procedure proposed in (Cardoso et al., 2010) to evaluate the stability of clustering solutions. The use of a weighted sample overcomes the need for selecting a classifier when performing cross-validation. Furthermore, sample dimension is not a severe limitation for implementing clustering stability evaluation, since the Indices of agreement values are based on the entire (weighted) sample, and not in a holdout sample – (Amorim and Cardoso, 2015b).

Studies on clustering stability summarized in Table 1 point out some issues. For example:

- "The stability of the clustering solution varies with the number of clusters that are inferred." - (Lange et al., 2004: 1302);

- Cluster structures can be recovered although no natural clusters exist in the data – e.g. (Dolnicar and Leisch, 2010; Steinley and Brusco, 2011). In particular, "data structure (other than density clusters) can cause some number of clusters to produce more reproducible results than others." - (Dolnicar and Leisch, 2010: 100). For example, ellipse data consisting of only one natural segment can be split into four segments in a stable manner.

In fact, despite the popularity of stability methods, its link to the underlying issue of clustering validity is still not well understood – e.g. (Shamir and Tishby, 2010; Dudoit and Fridlyand, 2002).

Luxburg (2009) discuss a series of theoretical results on clustering stability, namely the results obtained in (Ben-David et al., 2006; Ben-David et al., 2007; Ben-David and Luxburg, 2008; Shamir and Tishby, 2008a; Shamir and Tishby, 2008b) and in (Shamir and Tishby, 2010). Luxburg concludes that "While stability is relatively well-studied for the K-means algorithm, there does not exist much work on the stability of completely different clustering mechanisms." (p. 271). In fact, (Ben-David et al., 2006) conclude that a clustering algorithm is stable if its objective function has a unique global minimizer. These authors also point out the (sub)sampling process as a source of instability and state that the instability decreases as sample sizes (m) grow. (Ben-David and Luxburg, 2008) contest the use of stability especially for large samples and end up concluding that "stability is useful in one respect: high **instability can be used as an alarm sign** to distrust the clustering result, be it for sampling, algorithmic or other reasons." (p. 389). However, (Shamir and Tishby, 2008) show that even when the sample size increases the "convergence rate of clustering instability" may depend on K values used, the true K originating a faster convergence. These authors suggest the use the clustering distance (proportion of observations which are grouped into different clusters) scaled by $\sqrt{m}$, which rate of convergence dependent of K and is independent of the sample size used.

Indices of Agreement between partitions

Several similarity indices can be used to measure the agreement between two partitions of the same data - P^K and P^Q with K and Q groups, respectively - e.g. (Gower and Legendre, 1986; Milligan and Cooper, 1986). These indices are used in various domains namely, cluster validation, metaclustering, and consensus clustering - e.g. see (Jain et al., 1999; Krey et al., 2014).

A typology

When trying to measure similarity between two partitions one can resort to indices of association or simple agreement (ISA) or to indices of paired agreement (IPA), e.g. see (Cardoso, 2007; Warrens, 2008a). They are generally derived from the corresponding contingency table or cross-classification table $[n_{kq}]$ (corresponding to P^K and P^Q) where n_{k+} and n_{+q} refer to the total row and column counts, respectively.

Indices of simple agreement are, essentially, measures of association between two nominal variables which indicate the two partitions and are based on the number of observation that both partitions integrate in the same group (or not) – ISA examples are presented in Table 3.

Table 3 - Indices of simple agreement

ISA		Formula	Reference
NMI1	Normalized mutual information	$\frac{I(P^K,P^Q)}{Min\{H(P^K),H(P^Q)\}}$	(Strehl and Gohosh, 2002)
NMI2	Normalized mutual information	$\frac{I(P^K,P^Q)}{\sqrt{H(P^K)H(P^Q)}}$	(Strehl and Gohosh, 2002)
NMI3	Normalized mutual information	$\frac{2I(P^K,P^Q)}{H(P^K)+H(P^Q)}$	(Fred and Jain, 2003)
MIH	Mutual information ratio	$MI(P^K,P^Q)/H(P^K,P^Q)$	((Horibe, 1985), ((Kraskov et al., 2005))
NVI	Variation of information	$[H(P^K)-I(P^K,P^Q)]+[H(P^Q)-I(P^K,P^Q)]$	(Meilã, 2007)

The specific ISA examples considered (Table 3) are Mutual Information indices based in the concept of entropy. The entropy of a partition $H(P^K)$ quantifies the uncertainty about which cluster an individual randomly selected from the data set belongs to.

$$H(P^K) = -\sum_{k=1}^{K} \frac{n_{k+}}{n} log\left(\frac{n_{k+}}{n}\right) \quad (1)$$

and Mutual information is defined by:

$$I(P^K, P^Q) = \sum_{k=1}^{K}\sum_{q=1}^{Q}\frac{n_{kq}}{n} log\left({n_{kq}} \middle/ \frac{n_{k+}n_{+q}}{n} \right) \tag{2}$$

It quantifies the information shared between two partitions, of the same data set. Alternative normalizations may be considered for $I(P^K, P^Q)$ as illustrated in Table 3. Note that the $H(P^K, P^Q)$ referred in MIH is the entropy corresponding to the joint distribution of P^K and P^Q.

Indices of paired agreement are based on the number of pairs of observations that both partitions allocate (or not) to the same cluster and can also be written based on the elements of a similarity matrix (see Table 4). IPA examples are shown in Table 5.

Table 4 - Similarity matrix (counts of pairs of observations)

Partition P^K	**Partition P^Q**	
	Pair in same cluster	Pair not in same cluster
Pair in same cluster	a_{11}	a_{10}
Pair not in same cluster	a_{01}	a_{00}

Finally, it is worthwhile to note that the indices of agreement between two partitions can be used not only to evaluate the stability of clustering results - resorting to diverse (sub)sampling schemes as illustrated in the literature review (Table 1) - but also to evaluate their validity, comparisons being referred to *a priori* known classes.

The adjustment of indices for excluding agreement by chance

Consider the cross-classification Table 6 a). The agreement between partions as sets of disjoint classes $\{C_1, C_2\}$, and $\{C'_1, C'_2\}$ can be quantified resorting, for example, to the Rand index that equals 0.525. However, under the H$_0$ hypothesis of agreement by chance (assuming that the 2 partitions are constituted at random) the counts in the cross-classification table can be obtained by $n_{ij} = (n_{i+}n_{+j})/n$ - see Table 6 b). In this situation the Rand index value is 0.494, which is only slightly below the previously obtained value. This fact makes us wonder if the agreement quantified in Table 6 a).

Table 5 - Indices of paired agreement

IPA		Formula	Reference
R	Rand	$\frac{a_{11}+a_{00}}{a_{11}+a_{10}+a_{01}+a_{00}}$	(Rand, 1971)
GL	Gower & Legendre	$\frac{a_{11}+a_{00}}{a_{11}+0.5(a_{10}+a_{01})+a_{00}}$	(Gower and Legendre, 1986)
J	Jaccard	$\frac{a_{11}}{a_{11}+a_{10}+a_{01}}$	(Jaccard, 1908)
C	Czekanowski	$\frac{2a_{11}}{2a_{11}+a_{10}+a_{01}}$	(Czekanowski, 1932)
GK	Goodman & Kruskal	$\frac{a_{11}a_{00}-a_{10}a_{01}}{a_{11}a_{00}+a_{10}a_{01}}$	(Goodman and Kruskal, 1954)
SoS	Sokal & Sneath	$\frac{a_{11}a_{00}}{\sqrt{(a_{11}+a_{10})(a_{11}+a_{01})(a_{00}+a_{10})(a_{00}+a_{01})}}$	(Sokal and Sneath, 1963)
SS2	Sokal & Sneath (2)	$\frac{a_{11}}{a_{11}+2(a_{10}+a_{01})}$	(Sokal and Sneath, 1963)
FM	Fowlkes & Mallows	$\frac{a_{11}}{\sqrt{(a_{11}+a_{10})(a_{11}+a_{01})}}$	(Fowlkes and Mallows, 1983)

relates to a "ground truth" discovered by both partitions or it is due to chance. One should then have means to exclude the chance agreement from the Rand value. Using all the possible tables with identical marginal totals - see Table 11 - one easily finds the average Rand value (0.500).

Table 6 – a) Cross-classification data between two binary partitions (left)
b) Cross-classification data between two binary partitions constituted at random (right)

	C'_1	C'_2	Totals
C_1	30	20	50
C_2	10	20	30
Totals	40	40	80

	C'_1	C'_2	Totals
C_1	25	25	50
C_2	15	15	30
Totals	40	40	80

The subtraction of the Rand from the average 0.500, followed by a convenient normalization yields the adjusted Rand index value – 0.051 – which is obtained replacing *IA* by Rand in formula (3). It quantifies the relative agreement increase compared to agreement by chance.

$$IA_a(P^K,P^Q)=\frac{IA(P^K,P^Q)-E_{H0}[IA(P^K,P^Q)]}{Max[IA(P^K,P^Q)]-E_{H0}[IA(P^K,P^Q)]} \tag{3}$$

The fact that, in the Table 6 example the number of cross-classification tables with identical marginals is treatable, turns the calculation of $IA_a(P^K,P^Q)$ easy and extendable to various indices – see Table 7. However, for realistic problems the dimension of cross-classification tables undermines this method of calculation.

Several contributions are known to address the adjustment of indices of agreement which can be roughly categorized into: distributional approaches (DIST) and approximation methods (APPROX). The works on distributional approaches for the IPA indices can be viewed within the $\mathcal{L}$ family framework proposed by (Albatineh et al., 2006; Warrens, 2008; Warrens, 2008a). For example, the Rand and C indices adjustment is presented in Appendix C p.87. For the ISA mutual information index, the expected value, under H_0, is obtained by (Vinh et al., 2010):

$$E_{H_0}(MI)=\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{n_{kq}=max\{n_{k+}+n_{+q}-n,0\}}^{min\{n_{k+},n_{+q}\}}\frac{n_{kq}}{n}log\left(\frac{n\times n_{kq}}{n_{k+}\times n_{+q}}\right) \times\frac{n_{k+}!\,n_{+q}!\,(n-n_{k+})!\,(n-n_{+q})!}{n!\,n_{kq}!\,(n_{k+}-n_{kq})!\,(n_{+q}-n_{kq})!\,(n-n_{+q}-n_{k+}+n_{kq})!} \tag{4}$$

Alternative approximation approaches are suggested in (Albatineh and Niewiadomska-Bugaj, 2011) resorting to regression models - (Amorim and Cardoso, 2015c). The results from DIST and APPROX methods referred to the data example at hand are presented in Table 7. The missing cells illustrate the lack of means to calculate the adjusted values of several indices of agreement.

Table 7 – Indices average and adjusted values, under H_0, obtained by different methods

IA	**Averaged values**			**IA observed**	**Adjusted values**		
	TABLES LIST	**DIST**	**APPROX**		**TABLES LIST**	**DIST**	**APPROX**
Rand	0.500	0.500		0.525	0.051	0.051	
GL	0.666		0.668	0.689	0.067		0.063
J	0.341		0.340	0.364	0.035		0.036
C	0.509	0.509		0.534	0.293	0.051	
SoS	0.250			0.276	0.035		
NMI1	0.010	0.01		0.051	0.042	0.042	

The proposed approach

Experiments on stability

In this work we focus on the role of clustering stability in the evaluation of a clustering solution. For this end, we design an experiment based on synthetic data sets, with a known clustering structure, generated under 18 different scenarios.

In order to exclude agreement by chance when measuring the accordance between two partitions or crisp clustering solutions, a new method is proposed – IADJUST – that resorts to a simulation approach- (Amorim and Cardoso, 2015c).

The experimental analysis addresses:

1. The comparison between the stability of two clustering algorithms: an Expectation-Maximization algorithm (Rmixmod package (Lebret et al, 2012) for the estimation of a general Gaussian mixture model – $[P_K L_K C_K]$ in (Biernacki et al., 2006)) and a K-means algorithm ((Hartigan, 1975) algorithm implemented in the IBM SPSS Statistics software.)
2. The external validity of the clusterings obtained.

The IADJUST procedure

Given that in real examples is very difficult, or impossible, to obtain all tables, conditional on the row and column totals of the observed table (review Table 6 example), in the proposed method we resort to simulated cross-classification tables. The generation of tables is based on the Hypergeometric model. According to this model and given the values in the previous rows and columns, the conditional probability of the element n_{lm} of the cross-classification tables is given by the Hypergeometric distribution, e.g. (Patefield, 1981), with parameters: $\sum_{q=m}^{Q}\left(n_{+q} - \sum_{k=1}^{l-1} n_{kq}\right)$ (population size) $n_{l+} - \sum_{q=1}^{m-1} n_{+q}$ (number of successes in the population) and $n_{+m} - \sum_{k=1}^{l-1} n_{km}$ (sample size). Thus, the conditional expected value, under H_0, of n_{lm} given previous entries and the row and column totals can be calculated.

The IADJUST procedure begins by calculating the values of an IA, associated to the observed cross-classification table (observed indices values). Then, 17,000 tables are generated according the probabilistic method referred previously. For each generated table, the IA values are determined. These simulated IA values enable obtaining (under

H_0) the IA empirical distribution, the corresponding descriptive statistics (average, in particular) and the p-values estimates, i.e. the ratio between the number of simulated IA values greater than or equal to the IA observed value (see table 4 in (Amorim and Cardoso, 2015c)). According to Agresti et al. (1979) the 17,000 trials enabling to obtain estimates of the p-values with 99% confidence.

In IADJUST procedure is described in (Amorim and Cardoso, 2015b) - table 5). It is implemented in R language.

Testing IADJUST

In order to evaluate the precision of IADJUST we compare its results with the ones derived from analytical and approximation approaches - (Amorim and Cardoso, 2015c). The results referred to example in Table 6 a) are presented in Table 8.

Table 8 – Average and adjusted values obtained by IADJUST and with all possible tables

IA	Average values		IA observed	Adjusted values	
	IADJUST	TABLES LIST		IADJUST	TABLES LIST
Rand	0.49970	0.49968	0.5253	0.05121	0.05124
GL	0.66635	0.66634	0.6888	0.06726	0.06730
J	0.34144	0.34143	0.3644	0.03487	0.03490
C	0.50902	0.50900	0.5342	0.05121	0.05124
SoS	0.24975	0.24973	0.2760	0.03496	0.03498
NMI1	0.00967	0.00965	0.0511	0.04186	0.04188

IADUST into practice

We use adjusted IA to measure the stability of clustering solutions in diverse scenarios and better understand the relationship with external validity. We also capitalize on the results from IADJUST to obtain new insights concerning the distributions of diverse IA under H_0.

Simulated data sets

Experiments on stability rely on 540 simulated data sets. Design factors considered are the number of clusters, their balance and overlap (Table 9).

The increasing number of clusters is associated with increasing number of variables (2, 3 and 4 latent groups with 2, 3 and 4 Gaussian distributed variables) and, in order to deal with this increasing complexity, we consider data sets with 500, 800 and1100 observations, respectively.

We consider balanced settings (classes with similar dimensions) and unbalanced settings (classes with different a priori probabilities or weights).

Based on the measure of overlap between cluster adopted, (Maitra and Melnykov, 2010), we consider experimental scenarios with poorly separated, moderately separated and well separated clusters.

Table 9 - Simulated data sets for the experimental design.

K	Number of latent groups	n	Weights		Average overlap		
			Balanced	Unbalanced	poorly separated	moderately separated	well separated
2	2	500	0.5 , 0.5	0.3 , 0.7	0.6	0.15	0.02
3	3	800	0.3 , 0.3 , 0.4	0.6 , 03 0.1			
4	4	1100	0.25 , 0.25 0.25 , 0.25	0.5 , 0.25 0.15 , 0.10			

In order to generate the datasets within the scenarios, we capitalize on the contribution in (Maitra and Melnykov, 2010) and use the R MixSim package to generate structured data according to a finite Gaussian mixture model.

Real data sets

Experiments are also made resorting to real data with an *a priori* known clustering structure. The data sets used are found in the UCI Machine Learning Repository (Bache and Lichman, 2013) and summarized in Table 10.

Table 10 – Real data sets.

Data set	n	Features	Classes	Overlapping
Liver Disorders	345	6	C1 (145) C2 (200)	0.016
Wholesales	440	6	C1 (298) C2 (142)	0.111
Iris	150	4	Setosa (50) Versicolor (50) Virginica (50)	0.518
Wine recognition data	178	12	C1 (59) C2 (71) C3 (48)	0.002
Cars Silhouette	846	18	Bus (218) Saab (217) Opel (212) Van (199)	0.044
User Modeling	258	5	Very-low (24) Low (83) Middle (88) High (63)	0.028

A thesis Guide

The thesis chapters refer to the publications that are the result of the research conducted. The first publication “Clustering Cross-validation and Mutual Information Indices” applies the proposed IADJUST method to four Mutual Information indices in order to evaluate the agreement of clustering solutions with ground truth and also their stability. Based on this work we tuned the experimental scenarios to consider. The second publication “Paired Indices for Clustering Evaluation - Correction for Agreement by Chance” applies IADJUST to eight paired indices of agreement and goes one step further in evaluating the comparative performance of IADJUST with analytical and approximation approaches. The third paper “Comparing clustering solutions: the use of adjusted paired indices” capitalizes on the experimental work of the previous publication and adds theoretical support. It also produces a set of interesting results on the comparative performance of each index within experimental scenarios and also on the distributions of indices under the assumption of agreement by chance. Finally, the paper “Clustering stability and ground truth: numerical experiments” (Proceedings and

Paper publications) gathers the simple and paired-agreement views to evaluate the relationship between the clustering solutions agreement with ground truth and their stability. A Mutual Information index (MIH in Table 3) and the Adjusted Rand index provide new insights into this relationship within and inter scenarios.

CHAPTER 2: CLUSTERING CROSS-VALIDATION AND MUTUAL INFORMATION INDICES

This manuscript has the following reference:

Amorim, M. J. P. C. & Cardoso, M. G. M. S. Clustering cross-validation and mutual information indices. In: Ana Colubi, K. F., Gil Gonzalez-Rodriguez And Erricos JOhn Kontoghiorghes, ed. 20th International Conference on Computational Statistics (COMPSTAT 2012), 2012, Limassol, Cyprus. The International Statistical Institute / International Association for Statistical Computing, 39 -52.

Clustering cross-validation and mutual information indices

Maria J. Amorim, *Instituto Universitario de Lisboa (ISCTE – IUL), BRU–IUL, Lisboa, Portugal*, mjamorim@dmat.isel.pt
Margarida G. M. S. Cardoso, *ISCTE – Lisbon University Institute*, mgs@iscte.pt

Abstract. The use of cross-validation enables to evaluate clustering solutions internal stability. In this context mutual information indices (relying on the concept of entropy), enable to measure the agreement between diverse clustering solutions (partitions, in particular) based on diverse sub-samples drawn from the same source. In order to adequately do so, the proposed approach uses indices values corrected for agreement by chance. The corrected values are obtained using simulated indices values, corresponding to cross-classification tables generated under the hypothesis of restricted independence. The simulated tables have fixed marginal totals that match the ones derived from cross-validation. The proposed method is illustrated using real data referring to a retail application (with unknown structure) and using simulated data, with known structure with diverse clusters' weights and diverse degrees of cluster's overlap being considered. Furthermore, for simulated data, the agreement between clustering solutions obtained and the real partitions is also analyzed, enabling to discuss the observed relationship between stability and agreement with ground truth.

Keywords. Correction for chance agreement, Comparing partitions, Mutual information based indices, Stability.

1 Introduction

Clustering evaluation generally relies on some desirable properties of clustering solutions (partitions, in particular): the properties of clusters' compactness and separation, as well as the property of stability are often considered as indicators of clustering quality. In fact, since the real clustering is unknown (clustering being originated by an unsupervised process), one should focus on obtaining good enough partitions. The clustering solution's reproducibility in several data sets drawn from the same source may be used as an indicator of stability. In this setting, the comparison of alternative partitions based on slightly modified data sets may recur to a cross-validation procedure and mutual information indices may be used to summarize the agreement between partitions. In order to adequately evaluate the degree of agreement between partitions the entropy based indices are corrected to exclude agreement by chance using simulated indices of agreement, corresponding to cross-classification tables generated under the hypothesis of restricted independence to correct the observed indices values. This procedure is implemented with R software. Eight simulated data bases – normal mixtures exhibiting different degrees of overlap and different group proportions – are analysed. The evaluation of clustering solutions stability is complemented with the evaluation of the degree of agreement with the simulated data known structure. The clustering analysis of customers of a wholesale distributor complements the illustration of the proposed approach.

2 Methodological approach

The proposed method step–by–step

In order to evaluate the stability of a clustering solution and the degree of agreement with the simulated data known structure, this work focuses on the performance of mutual information based indices to determine the agreement between partitions being compared. The experimental work is delineated as follows:

1. A data set is considered for analysis;

2. Clustering analysis (discovering a partition) is performed recurring to two alternative algorithms that predictably will capture the latent structure in the data set with different levels of precision and provide solutions that will evidence diverse degrees of stability;

3. Stability of the clustering solutions is evaluated using a cross–validation procedure yielding 4 entropy based measures that express the same stability;

4. Entropy based measures of agreement are corrected for agreement by chance in an attempt to capture the essential stability. For that end, the proposed method determines indices thresholds which are based on the simulation of contingency tables emulating agreement by chance between clustering solutions;

5. The evaluation of stability takes place using normalized mutual information indices values corrected by chance;

6. Finally, in the case of simulated data sets, the agreement with the real structure is also measured with corrected indices and the correlation between this agreement and stability is analyzed.

Eight experimental data sets are considered, each with 800 observations, 3 latent groups and 3 conditionally independent Gaussian distributed variables. Two types of data sets are generated: balanced (clusters having similar weights) and disproportionate (clusters with different weights). Balanced mixture clusters' weights are 0.40, 0.30 and 0.30. Disproportionate mixtures clusters' weights are 0.60, 0.30 and 0.10. The generated mixtures also exhibit diverse degrees of cluster's overlapping being: A) very–poorly–separated, B) poorly–separated, C) moderately–separated and D) very–well–separated. The *a priori* established degree of overlap is the sum of misclassification probabilities and the R MixSim package is used to obtain the simulated data [11]. In order to obtain two alternative clustering solutions (partitions) and compare the correspondent stability, a model-based clustering method (Expectation – Maximization based algorithm) and a (dis)similarity-based clustering method (a K-means algorithm) are used [7]. For simulated data sets the true number of groups is an input both in the K-Means and EM based procedures. Cross-validation enables to test the clustering solutions stability comparing partitions originating from different data sub-samples. In fact, cross-validation is commonly known in supervised data analysis but rarely used in the unsupervised domain. This concept (of cross-validation) was firstly imported from supervised to unsupervised analysis by [5] and used specifically in clustering applications since then (e.g. [19] uses cross-validations to determine the number of clusters). A variant of the usual cross-validation procedure–weighted cross-validation [6] – is used in the present work. This approach relies on weighted samples and calculates the indices of agreement between partitions based on the entire weighted sample and overcomes the need to build supervised classifiers to export the clustering from the training to the test sample, significantly simplifying the original [5] procedure as well as similar variants (e.g. [12]). The result from cross-validation is a contingency table that summarizes the agreement between a partition derived from a weighted training sample (unit weights are considered for half of the sample and quasi-zero weights – 10^{-10} – for the remaining half) and a partition relying on a weighted test sample (unit weights for the test observations and quasi-zero weights for the training observations). Diverse mutual information indices are computed for each cross-validation contingency table summarizing the agreement between the "training" and "test" partitions. These indices do not take into account the possibility of agreement by chance between the two partitions. In fact, the agreement between two partitions should exceed the agreement by chance. Hubert and Arabie, [9], acknowledged that and proposed the adjusted the Rand index which subsequently has shown to perform well in the context of clustering evaluation,[14]. Since then, several authors adopted similar approaches (e.g. [3] and [2]). In the line of these authors work, the present research addresses the correction of mutual information indices for agreement by chance. It recurs to the simulation of cross-classification tables having the same marginal totals as the contingency table yielded by cross-validation to provide the adequate indices thresholds, [4]. Finally, the performance of the diverse mutual indices can be compared.

Indices of agreement between partitions – the mutual information indices

There are several indices of agreement (IA) which can be used to compare partitions and ultimately determine their stability. In the present study, diverse mutual information indices are used as IA between partitions. They are based in the concept of entropy and can be determined from a cross-classification table between the two partitions (P^K and P^Q with K and Q groups) being considered. The cross-classification table is then a KxQ matrix, whose $(k,q)^{th}$ element $-n_{kq}-$ is the number of observations in the intersection of clusters C_k of P^K and C_q of P^Q. Let $n_{k.}$ and $n_{.q}$ represent the table's row and column totals and n the number of observation, respectively. The entropy of a partition $-H(P^K)-$ quantifies the uncertainty about which cluster an individual random selected from the data set belongs to:

$$H(P^K) = -\sum_{k=1}^{K} \frac{n_{k.}}{n} \log(\frac{n_{k.}}{n}). \tag{1}$$

In the clustering context, mutual information $I(P^K, P^Q)$ measures the information shared between two partitions of the same data set:

$$I(P^K, P^Q) = \sum_{k=1}^{K}\sum_{q=1}^{Q} \frac{n_{kq}}{n} \log(\frac{n_{kq}}{\frac{n_{k.}n_{.q}}{n}}). \tag{2}$$

Mutual information is a metric on the space of all partitions of the same data set. The Variation of information $VI(P^K, P_Q)$ proposed by [13] is also based on mutual information and measures the distance between two partitions of the same data set being also a metric.

$$\begin{aligned} VI(P^K, P^Q) &= H(P^K) + H(P^Q) - 2I(P^K, P^Q) \\ &= [H(P^K) - I(P^K, P^Q)] + [H(P^Q) - I(P^K, P^Q)]. \quad (3) \\ & \quad (4) \end{aligned}$$

Mutual information and Variation of Information are used as IA between partitions to evaluate their stability. In order to have the indices values bounded between 0 and 1 normalized versions are used: NMI1, NMI2 and NMI3 for the mutual information indices and NVI, a normalized VI (see [17], [8] and [13]). Since

$$I(P^K, P^Q) \leq Min\{H(P^K), H(P^Q)\}, \tag{5}$$

and

$$VI(P^K, P^Q) \leq \log(n), \tag{6}$$

the following normalized mutual information indices are determined:

$$\begin{aligned} NMI1(P^K, P^Q) &= \frac{I(P^K, P^Q)}{Min\{H(P^K), H(P^Q)\}}, \\ NMI2(P^K, P^Q) &= \frac{I(P^K, P^Q)}{\sqrt{H(P^K)H(P^Q)}}, \\ NMI3(P^K, P^Q) &= \frac{2I(P^K, P^Q)}{H(P^K) + H(P^Q)}, \\ NVI(P^K, P^Q) &= \frac{VI(P^K, P^Q))}{\log(n)}. \end{aligned} \tag{7}$$

Finally, since NMI1, NMI2 and NMI3 are inversely correlated with NVI, 1-NVI is used.

Indices corrected by chance – the use of thresholds

IA values are able to measure the agreement between partitions drawn from slightly modified data sets to decide upon a clustering solution stability. However, the agreement between two partitions can be due to chance. The first proposal to consider indices of agreement (IA) between partitions corrected by chance is due to Hubert and Arabie, [9]. For correction, they consider the mean of the Rand index [16] under the assumption of restricted independence – H_0 – and rely on a generalized hypergeometric distribution referred to the entries of a cross-classification table with fixed marginal totals to obtain the expected value or threshold, Mean(IA). The adjusted index is then:

	Database	A		B		C		D	
Group	Variable	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
1(30%)	X_1	10.30	2.80	10.50	3.50	11.90	1.10	10.50	1.00
	X_2	4.80	0.80	2.30	0.50	2.50	0.30	2.50	1.30
	X_3	7.20	2.00	7.80	2.00	8.00	0.90	4.30	1.80
2(30%)	X_1	10.00	2.60	10.00	3.00	9.80	1.20	15.00	2.20
	X_2	4.50	0.50	2.50	0.30	1.50	0.30	4.00	1.20
	X_3	7.00	1.40	7.00	1.00	6.80	0.70	7.00	1.50
3(40%)	X_1	9.50	2.20	9.50	2.00	11.80	1.40	7.00	2.30
	X_2	4.30	0.40	2.00	0.40	2.00	0.40	6.20	1.60
	X_3	7.40	1.50	7.50	1.20	8.90	0.70	2.50	1.70
Average	overlap	0.739	-	0.633	-	0.140	-	0.019	-
Maximum	overlap	0.800	-	0.653	-	0.516	-	0.029	-

Table 1: The Gaussian distributed simulated balanced data sets

$$adj_M(IA) = \frac{IA_{obs} - Mean(IA)}{1 - Mean(IA)}. \tag{8}$$

The adjusted or corrected index is bounded by 1 and takes the value 0 when the observed IA – IA_{obs} – is the expected value.

To calculate the exact IA expected value under H_0 one would have to consider all the cross-classification tables with observed marginal totals. This is only feasible for relatively small tables with relatively small observed counts due to computational complexity [10]. In the present work, the expected value of each mutual information based index is estimated using the average of its values corresponding to cross-classification tables generated under H_0 – see [4]. The tables generated have the same marginal totals as the cross-validation resulting table. For each generated table, the IA values are determined which enable obtaining the empirical IA distribution under H_0 and the corresponding descriptive statistics. The p-values are estimated as the ratio between the number of simulated IA values greater than or equal to the IA observed value and 17000 (17000 trials enabling to obtain estimates with 99 percent confidence [1]).

The proposed approach yields each IA empirical distribution enabling to determine diverse thresholds: the average threshold value and the 99 percentile threshold are used in this work.

3 Data analysis and results

Simulated data sets

The simulated data sets are A) very-poorly-separated, B) poorly-separated, C) moderately-separated and D) very-well-separated with the distributional parameters presented in Tables 1 and 2.

Stability results obtained for data sets A are summarized in Tables 3 and 5. Since the average values and 99 percentile IA values are very similar (for all data sets), only the Mean corrected values are reported for data sets B, C and D (see Tables 4 and 6). Stability results are summarised in Figure 1.

The results regarding the agreement with the original clustering structure are reported in Tables 7, 8 and in Figure 2 improving with the decrease of overlap (as expected) and the performance of EM clearly surpassing the performance of KM. The NMI indices provide very similar values but the 1-NVI yields much higher values and a lower variation range. However, after correction, they all perform similarly.

When the EM algorithm is used, the obtained IA values in the analysis of the agreement with the real data structure and in the analysis of the stability exhibit a high positive linear correlation. For example, the NMI1 values have a Pearson correlation coefficient equal to 0.999 for balanced data sets and equal to 0.997 for disproportionate data sets. When the KM algorithm is used the IA values have a moderate positive correlation, for example, the NMI1 values have a Pearson correlation coefficient equal to 0.617 for balanced data sets and equal to 0.524 for disproportionate data sets.

Real data set - segmenting clusters of a wholesale distributor

Real data referring to 440 customers of a wholesale distributor - 298 from the Horeca (Hotel/Restaurant/Café) channel and 142 from the Retail channel - are considered for segmentation.

The wholesale data includes the annual spending in monetary units (m.u.) on diverse product categories, namely: fresh products, milk products, grocery, frozen products, detergents and paper products and delicatessen. Since the EM solution is more stable than KM solution (results

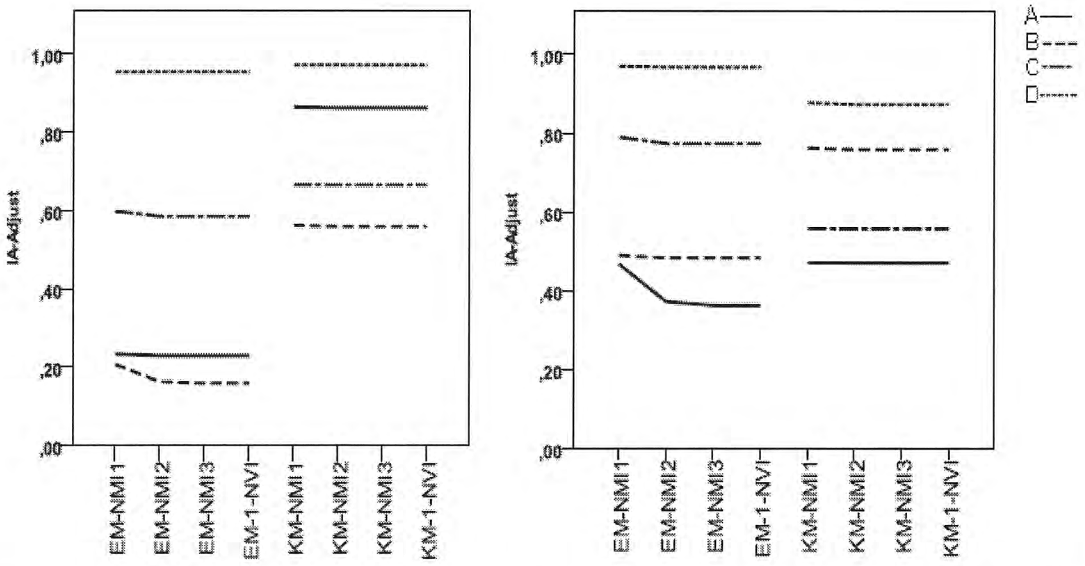

Figure 1: Stability results – balanced data sets (on the left) and disproportionate data sets (on the right)

	Database	A		B		C		D	
Group	Variable	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
1(60%)	X_1	10.80	2.80	11.00	2.20	12.30	1.10	14.30	0.70
	X_2	5.10	0.80	5.30	0.80	6.40	0.60	7.00	0.20
	X_3	7.40	2.00	7.80	1.80	8.80	1.10	9.20	0.30
2(30%)	X_1	10.00	2.60	10.00	2.00	11.00	1.00	12.70	0.50
	X_2	4.50	0.50	4.50	0.50	5.00	0.50	5.00	0.40
	X_3	7.00	1.40	7.20	1.40	7.50	0.80	7.60	0.30
3(10%)	X_1	9.20	2.00	9.40	1.80	9.50	0.90	11.00	0.50
	X_2	4.10	0.40	4.00	0.40	3.70	0.40	3.50	0.30
	X_3	6.90	1.50	7.00	1.50	6.60	0.70	6.00	0.20
Average	overlap	0.758	-	0.632	-	0.143	-	0.021	-
Maximum	overlap	0.917	-	0.866	-	0.215	-	0.115	-

Table 2: The Gaussian distributed simulated disproportionate data:

	Dataset			A			
	IA	value	p-value	P_{99}	adj_{P99}	mean	adj_M
EM	NMI1	0.236	0.000	0.009	0.229	0.003	0.234
	NMI2	0.231	0.000	0.009	0.224	0.003	0.229
	NMI3	0.231	0.000	0.009	0.224	0.003	0.229
	1-NVI	0.782	0.000	0.720	0.224	0.718	0.229
KM	NMI1	0.865	0.000	0.008	0.863	0.002	0.864
	NMI2	0.862	0.000	0.008	0.861	0.002	0.862
	NMI3	0.862	0.000	0.008	0.861	0.002	0.862
	1-NVI	0.956	0.000	0.680	0.861	0.678	0.862

Table 3: Stability results for balanced data set A

	Dataset	B		C		D	
	IA	value	adj_M	value	adj_M	value	adj_M
	NMI1	0.211	0.207	0.600	0.599	0.954	0.954
EM	NMI2	0.166	0.163	0.587	0.586	0.954	0.954
	NMI3	0.161	0.159	0.586	0.585	0.954	0.954
	1-NVI	0.794	0.159	0.867	0.585	0.985	0.954
	NMI1	0.563	0.562	0.667	0.666	0.972	0.972
KM	NMI2	0.560	0.559	0.666	0.665	0.972	0.972
	NMI3	0.560	0.559	0.666	0.665	0.972	0.972
	1-NVI	0.857	0.559	0.890	0.665	0.991	0.972

Table 4: Stability results for balanced data sets: B, C and D

	Dataset			A			
	IA	value	p-value	P_{99}	adj_{P99}	mean	adj_M
	NMI1	0.471	0.000	0.022	0.459	0.006	0.468
EM	NMI2	0.376	0.000	0.017	0.365	0.005	0.373
	NMI3	0.366	0.000	0.017	0.355	0.005	0.363
	1-NVI	0.921	0.000	0.878	0.355	0.876	0.363
	NMI1	0.472	0.000	0.008	0.468	0.002	0.471
KM	NMI2	0.472	0.000	0.008	0.468	0.002	0.471
	NMI3	0.472	0.000	0.008	0.468	0.002	0.471
	1-NVI	0.829	0.000	0.679	0.468	0.678	0.471

Table 5: Stability results for disproportionate data set A

obtained are summarized in Table 9) only the main descriptive measures of the EM solution are presented in Table 10. There is a segment 3) with high expenditures. Segments 1) and 2) refer mainly to Horeca and retail clients, respectively. The segments are very–well–separated, EM solution has a average overlap equal to 0.016 and a maximum overlap equal to 0.033.

4 Discussion and perspectives

Overall, the comparison of clustering stability relying on the adjusted indices tends to agree with the conclusions derived from the indices with no correction for chance. One clear advantage of the proposed approach is to enable the comparison between the performance of multiple indices which may exhibit very different values and distributions and brings some new insights. E.g. the correction for chance is particularly relevant when considering poorly–separated clusters namely for the 1–NVI index. After correction, the IA values present similar values. In general, the IA

	Dataset	B		C		D	
	IA	value	adj_M	value	adj_M	value	adj_M
EM	NMI1	0.492	0.490	0.790	0.790	0.971	0.970
	NMI2	0.485	0.484	0.774	0.774	0.968	0.968
	NMI3	0.485	0.484	0.774	0.773	0.968	0.968
	1-NVI	0.878	0.484	0.938	0.773	0.992	0.968
KM	NMI1	0.763	0.762	0.559	0.558	0.878	0.878
	NMI2	0.759	0.758	0.558	0.557	0.873	0.873
	NMI3	0.759	0.758	0.558	0.557	0.873	0.873
	1-NVI	0.921	0.758	0.856	0.557	0.959	0.873

Table 6: Stability results for disproportionate data sets: B, C and D

	Dataset	A		B		C		D	
	IA	value	adj_M	value	adj_M	value	adj_M	value	adj_M
EM	NMI1	0.054	0.048	0.044	0.038	0.449	0.448	0.890	0.890
	NMI2	0.033	0.029	0.028	0.024	0.448	0.447	0.890	0.889
	NMI3	0.029	0.026	0.025	0.022	0.448	0.447	0.890	0.889
	1-NVI	0.784	0.026	0.780	0.022	0.819	0.447	0.964	0.889
KM	NMI1	0.016	0.013	0.021	0.019	0.384	0.383	0.888	0.888
	NMI2	0.016	0.013	0.021	0.019	0.383	0.382	0.888	0.887
	NMI3	0.016	0.013	0.021	0.019	0.383	0.382	0.888	0.887
	1-NVI	0.681	0.013	0.682	0.019	0.798	0.382	0.963	0.887

Table 7: Agreement with the original structures – balanced data sets

empirical distribution, under the assumption of agreement by chance, exhibits a positive asymmetry. This is valid in the application as well as for simulated data (see Figure 3). According to the obtained results for the balanced data sets, the KM clustering results exhibit more stability then the EM results. However, the EM solutions are more stable when referring to the disproportionate data sets. Furthermore, EM performs better when referring to agreement with the original structure. The question remains whether, in the case of poorly separated clusters, the latent structure can be properly be designated as a clustering structure. The partitions obtained for the balanced data sets show a better agreement with the real structure of the data than those obtained for the disproportionate data sets. The correlation between agreement with the real data structure and the stability of the solutions depends on the algorithm that is used, the EM solutions have a higher correlation than those obtained with KM. In the future, the performance of the mutual information indices at a condition regarding the number of clusters/number of variables/degree of overlapping and degree of balance as well as the trade-off bias stability should

	Dataset	A		B		C		D	
	IA	value	adj_M	value	adj_M	value	adj_M	value	adj_M
EM	NMI1	0.113	0.111	0.225	0.222	0.649	0.648	0.964	0.964
	NMI2	0.108	0.106	0.195	0.193	0.642	0.641	0.962	0.962
	NMI3	0.108	0.106	0.193	0.191	0.642	0.641	0.962	0.962
	1-NVI	0.744	0.106	0.815	0.191	0.900	0.641	0.990	0.962
KM	NMI1	0.053	0.050	0.105	0.103	0.472	0.470	0.644	0.643
	NMI2	0.049	0.046	0.094	0.092	0.439	0.437	0.582	0.581
	NMI3	0.049	0.046	0.094	0.091	0.438	0.436	0.579	0.577
	1-NVI	0.718	0.046	0.733	0.091	0.829	0.436	0.876	0.577

Table 8: Agreement with the original structures – disproportionate data sets

	IA	value	p-value	P_{99}	adj_{P99}	mean	adj_M
EM	NMI1	0.811	0.000	0.017	0.807	0.005	0.810
	NMI2	0.804	0.000	0.017	0.800	0.005	0.803
	NMI3	0..804	0.000	0.017	0.800	0.005	0.803
	1-NVI	0.941	0.000	0.702	0.800	0.699	0.803
KM	NMI1	0.323	0.000	0.025	0.306	0.008	0.318
	NMI2	0.302	0.000	0.023	0.286	0.008	0.297
	NMI3	0.302	0.000	0.023	0.285	0.008	0.296
	1-NVI	0.855	0.000	0.797	0.285	0.794	0.296

Table 9: Stability results for the Wholesale data set

Segments		Fresh	Milk	Grocery	Frozen	Det.Paper	Delicassen
1	mean	12968	2237	27773	3330	435	971
(54%)	std.dev.	11387	1602	1736	3382	375	854
2	mean	7506	7815	12167	1293	5073	1286
(37%)	std.deV.	7385	3969	6531	1213	3457	1016
3	mean	24736	19221	22167	8882	8772	5910
(9%)	std.dev.	23377	16261	19798	12078	10833	7812
data set	mean	12000	5796	7951	3072	2881	1525
	std.dev.	12647	7380	9503	4855	4768	2820

Table 10: Segments annual spending values (m.u.)

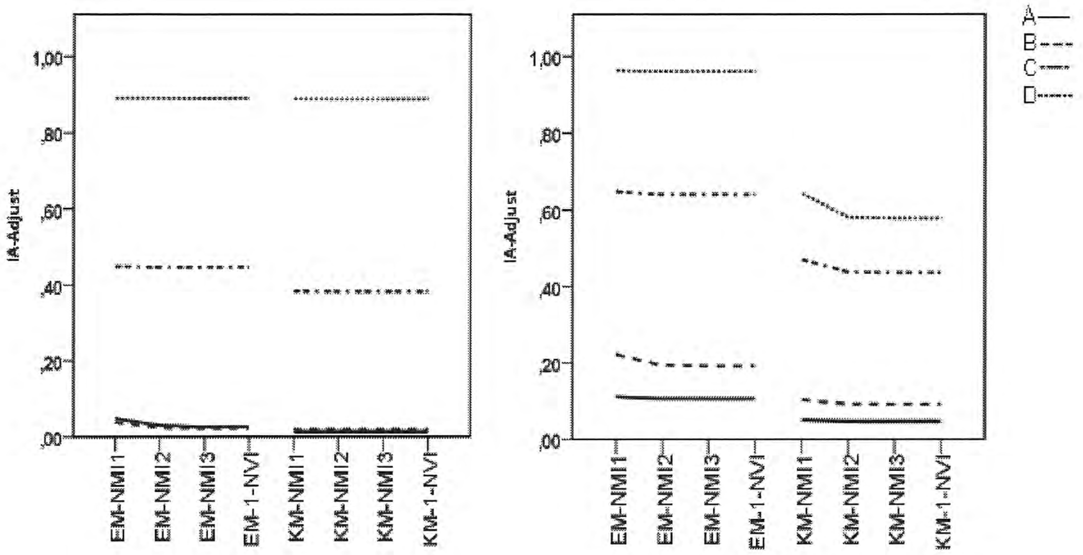

Figure 2: Agreement with the original structures – balanced data sets (on the left) and disproportionate data sets (on the right)

be further explored.

Bibliography

[1] Agresti, A., Wackerly, D. and Boyett, J. M. (1979) *Exact Conditional Tests for Cross – Classifications: Approximation of Attained Significance Levels.* Psychometrika. **44**, 75–83.

[2] Albatineh, Ahmed N. and Niewiadomska-Bugaj, Magdalena (2011) *MCS: A Method for Finding the Number of Cluesters.* Journal of Classification. **28** 184209.

[3] Albatineh, A. N. (2010) *Means and Variances for a Family of Similarity Indices Used in Cluster Analysis.* Journal of Statistical Planning and Inference. **140**, 2828–2838.

[4] Amorim, M. J. P. C. and Cardoso, M. G. M. S. (2010) *Limites de concordancia entre duas partições.* Livro de resumos, XVIII Congresso Anual da Sociedade Portuguesa de Estatística. **1**, 47–48.

[5] Breckenridge, J. N. (1989) *Replicating cluster analysis: Method, consistency, and validity.* Multivariate Behavioral Research. **24**, 147-161.

[6] Cardoso, M. G. M. S., Carvalho, A. P. L. and Faceli, K. (2009) *Evaluation of clustering results:the trade–off bias–variability. In Classification as a Tool for Research.* Proceedings of the 11th IFCS Biennial Conference. Dresden, March 13–18, 2009 Studies in Classification, Data Analysis, and Knowledge Organization Springer, Berlin-Heidelberg-New York, 201–208.

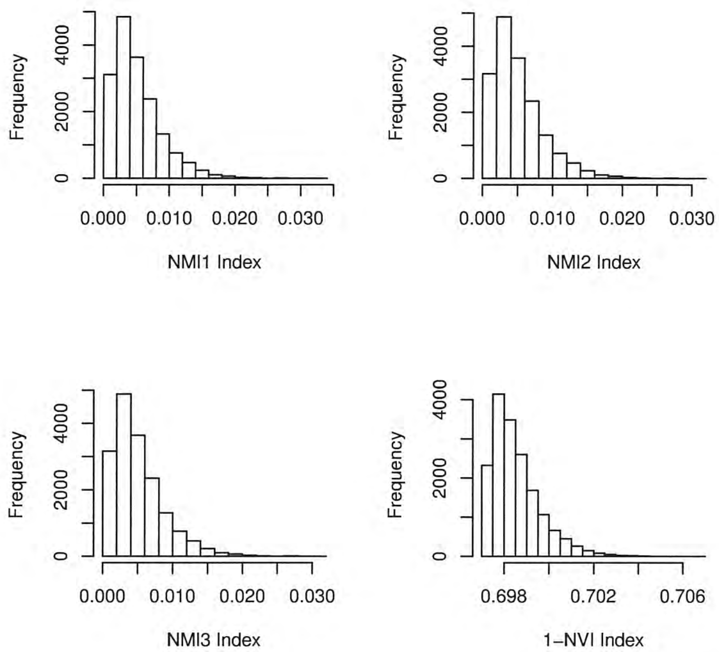

Figure 3: The IA empirical distribution (Wholesale data set, EM results)

[7] Everitt, B., Landau, S. and Morven, L. (2001) *Cluster Analysis.* 4th Ed. Arnold.

[8] Fred, A. and Jain,A. K. (2003) *Robust data clustering.* In Proceedings of the IEEE Computer Society Conference on Computer Vision Pattern Recognition. CVPR.

[9] Hubert , L. and Arabi, P. (1985) *Comparing Partitions.* Journal of Classification. **2**, 193–218.

[10] Krzanowski, W. J. and Marriott, F. H. C. (1994) *Multivariete Analysis Part 1 Distributions, ordination and inference.* Edward Arnold ed.

[11] Maitra, R. and Melnykov, V. (2010) *Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms.* Journal of Computational and Graphical Statistics. **2**, 354–376.

[12] McIntyre, R. M. and Blasheld, R. K. (1980) *A Nearest-Centroid Technique for Evaluating the Minimum – Variance Clustering Procedure.* Multivariate Behavioral Research. **15**, 225–236.

[13] Meila, M. (2007) *Comparing Clusterings – an information based distance.* Journal of Multivariate Analysis. **98**, 873–895.

[14] Milligan, G. W. and Cooper, M. C. (1986) *A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis.* Multivariate Behavioral Research. **21**, 441–458.

[15] Mirkin, B. (1996) *Mathematical Classification and Clustering.* Dordrecht/Boston/London. Kluwer Academic Plublishers.

[16] Rand, W. M. (1971) *Objective Criteria for the Evaluation of Clustering Methods.* Journal of the American Statistical Association. **66**, 846–850.

[17] Strehl, A. and Gohosh, J. (2002) *Cluster ensembles– a knowledge reuse framework for combining partitions.* Journal of Machine Learning Research, **3**, 583–617.

[18] Tibshirani, R., Walther, G., Botstein, D. and Brown, P. (2005) *Cluster validation by prediction strength.* Journal of Computational and Graphical Statistics. **14(3)** 511–528.

[19] Wang, J.(2010) *Consistent selection of the number of clusers via cross– validation.* Biometrika. **97** 893–904. 511–528.

CHAPTER 3: PAIRED INDICES FOR CLUSTERING EVALUATION - CORRECTION FOR AGREEMENT BY CHANCE

This manuscript has the following reference:

Amorim, M. J. & Cardoso, M. G. M. S. Paired Indices for clustering Evaluation. Correction for Agreement by Chance. In: The, I. P. O., ed. 16 th International Conference on Enterprise Information Systems, 2014 Lisboa, Portugal, 27-30 Abril. 164-170.

Paired Indices for Clustering Evaluation
Correction for Agreement by Chance

Maria José Amorim[1] and Margarida G. M. S. Cardoso[2]

[1]*Dep. of Mathematics, ISEL and Inst. Univ. de Lisboa (ISCTE-IUL), BRU-IUL, Av. Forças Armadas, Lisboa, Portugal.*
[2]*Dep. of Quantitative Methods and BRU-UNIDE, ISCTE Busines School-IUL, Av. das Forças Armadas, Lisboa, Portugal.*
mjamorim@adm.isel.pt, margarida.cardoso@iscte.pt

Keywords: Adjusted Indices, Indices of Paired Agreement, Clustering Evaluation.

Abstract: In the present paper we focus on the performance of clustering algorithms using indices of paired agreement to measure the accordance between clusters and an *a priori* known structure. We specifically propose a method to correct all indices considered for agreement by chance – the adjusted indices are meant to provide a realistic measure of clustering performance. The proposed method enables the correction of virtually any index – overcoming previous limitations known in the literature - and provides very precise results. We use simulated datasets under diverse scenarios and discuss the pertinence of our proposal which is particularly relevant when poorly separated clusters are considered. Finally we compare the performance of EM and K-Means algorithms, within each of the simulated scenarios and generally conclude that EM generally yields best results.

1 INTRODUCTION

In the present study we focus on the use of indices of paired agreement to measure accordance between two partitions of the same data and propose a method to handle agreement by chance.

This contribution aims to fill a gap in the literature since recent alternative solutions that have been proposed to address this issue - e.g. (Albatineh, 2010) or (Albatineh and Niewiadomska -Bugaj, 2011) - are limited in scope. We resort to diverse indices of paired agreement – Rand, Russell and Rao, Gower and Legendre, Jaccard, Czekanwski, Goodman and Kruskal, Sokal and Sneath, Fowlkes and Mallows – and illustrate the capacity of the proposed method to adjust virtually any index for agreement by chance.

In order to illustrate the usefulness of the proposed method we compare the performance of two well-known clustering tools: the Expectation Maximization (EM) and the K-Means (KM) algorithms. The EM provides the estimation of a finite mixture model - (Dempster et al., 1977) and, for example, (O'Hagan et al., 2012). The KM algorithm, a (dis)similarity-based clustering method, was independently discovered in different scientific fields and is still a widely used clustering tool ((Jain, 2010), (Shamir and Tishby, 2010)).

We conduct clustering external validation trying to measure the fit between a clustering structure captured in cluster analysis and the ground truth. The numerical experiments conducted resort to simulated data sets and consider diverse clustering scenarios.

1.1 Indices of Paired Agreement between Partitions

Similarity indices have been used in various domains for a long time: e.g. in clustering ecological species (Jaccard, 1908), in plant genetics (Meyeri et al., 2004) or in documents clustering (Chumwatana et al., 2010). Several similarity indices can be used to measure the agreement between two partitions of the same data - P^K and P^Q with K and Q groups, respectively. These are generally designated by Indices of Agreement (IA) - see ((Gower and Legendre, 1986), (Milligan and Cooper, 1986)).

Some of the IA are based on the number of pairs of observations that both partitions allocate (or not) to the same cluster – these are Indices of Paired Agreement (IPA). In the present study, diverse IPA are used to measure the degree of agreement between partitions. They can be determined from a similarity matrix **A** - a 2×2 matrix, where element a=A(1,1) represents the number of pairs of observations both partitions agree to allocate in the same group; b=A(1,2) represents the number of pairs that only belong to the same group in partition P^K; c=A(2,1) represents the numbers of pairs that only belong to the same group in partition P^Q; d=A(2,2) represents the number of pairs of observations both partitions agree to allocate to different groups. The values of a, b, c and d can be calculated from the cross-classification table between the two partitions being considered (see equations 1 to 4). The cross-classification table is a K*Q matrix, whose (k,q)th element - n_{kq} - is the number of observations in the intersection of group C^k of P^K (k=1...K) and C^q of P^Q(q=1...Q), $n_{k.}$ and $n_{.q}$ represent the matrix's rows and columns totals (respectively) and n the number of observations.

$$a = \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^2 - \frac{n}{2} \qquad (1)$$

$$b = \frac{1}{2}\sum_{q=1}^{Q} n_{.q}^2 - \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^2 \qquad (2)$$

$$c = \frac{1}{2}\sum_{k=1}^{K} n_{k.}^2 - \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^2 \qquad (3)$$

$$d = \binom{n}{2} - a - b - c \qquad (4)$$

Table 1: Indices of paired agreement.

IPA	ℒ Family	
R	✓	(Rand, 1971)
RR	✓	(Russell and Rao, 1940)
GL	×	(Gower and Legendre, 1986)
J	×	(Jaccard, 1908)
C	✓	(Czekanwski, 1932)
GK	×	(Goodman and Kruskal, 1954)
SoS	×	(Sokal and Sneath, 1963)
SS2	×	(Sokal and Sneath, 1963)
FM	✓	(Fowlkes and Mallows, 1983)

In the present work we consider the indices of paired agreement in Table 1. The indices SoS and SS2 can be calculated using the equations (5) and (6), respectively, the others indices equations can be found in references mentioned in Table 1.

$$SoS = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} \qquad (5)$$

$$SS2 = \frac{a}{a+2(b+c)} \qquad (6)$$

1.2 Correcting Indices for Agreement by Chance

In the context of clustering validation, indices of agreement (IA) are used to measure the agreement between partitions drawn from slightly modified data sets to decide upon a clustering solution stability, or to measure the agreement between clustering solutions and the real partition (external validation). The relevance of clustering validation is underlined by (Hennig, 2006), for example.

The agreement between two partitions – summarized in the corresponding cross-classification table –can, however, be due to chance. Therefore, in order to adequately evaluate the degree of agreement between two partitions, indices of agreement must be corrected to exclude agreement by change. (Hubert and Arabie, 1985) were the first to address this issue regarding the Rand index. For correction, they considered the mean of this paired index under the null hypothesis (H_0) of no association between the partitions to be compared, conditional on the row and column table totals - hypothesis of restricted independence. The adjusted index is then:

$$adj_M(IPA) = \frac{IPA_{obs} - Mean(IPA)}{1 - Mean(IPA)} \qquad (7)$$

Adj_M(IPA) is bounded by 1 and takes the value zero when the observed index - IPA $_{obs}$– is equal to the expected value under H_0.

In general, the exact IPA mean, under H_0, can be determined considering all the cross-classification tables under the hypothesis of restricted independence. However, this is only feasible for relatively small tables with small observed counts, due to computational complexity (Krzanowski and Marriott, 1994).

Under H_0, the probability of observing the associated cross-classification table can be modelled by the Multivariate Hypergeometric distribution (Halton, 1969) and the conditional probability of the value n_{kq} given the values in the previous rows and columns can be modelled by the Hypergeometric distribution. Thus, the conditional expected value of n_{kq} given previous entries and the row and column totals can be calculated under H_0. In fact, one can determine the means and variation of all IPA that are linear functions of the sum of the squares of the n_{kq} i.e. all indices belonging to the ℒ Family, namely the R, RR, C and FM indices in Table 1- see (Albatineh,

2010) for more details. (Albatineh and Niewiadomska-Bugaj, 2011) proposed an alternative approach for some indices - SS2, J and GL - that are not members of the $\mathcal{L}$ Family. They expressed J and SS2 as functions of C, and GL as function of R, and approximately computed their expected values.

Despite the diverse approaches to handle the correction for agreement by chance there are various IPA that are not covered by the procedures so far proposed - GK and SoS, for example. Therefore we propose a methodology that can deal with the correction of any IPA for agreement by chance.

2 THE PROPOSED METHOD

In the present work, the expected value of each IPA is estimated using the average of its values corresponding to 17,000 cross-classifications tables generated under H_0 - see (Amorim and Cardoso, 2010). For each generated table, the IPA values are determined which enables obtaining the empirical IPA distribution (under H_0) and the corresponding descriptive statistics.

The 17,000 cross-classifications tables generated ensure that average estimates have 99% confidence (Agresti et al., 1979).

The advantage of the proposed approach is that it can be applied to virtually all indices– see also (Amorim and Cardoso, 2012) where a similar procedure was used for Mutual Information Indices.

In order to evaluate the performance of the IPA in this study (seeTable 1), several scenarios are considered:

–Simulated data sets with Gaussian 2, 3 and 4 latent groups with 2, 3 or 4 Gaussian distributed variables and with 500, 800 and 1100 observations, respectively.

–Mixtures with balanced and unbalanced clusters' weights.

–Diverse degrees of clusters' overlapping: poorly-separated, moderately-separated and well-separated clusters, where the degree of overlap is the sum of misclassification probabilities (Maitra and Melnykov, 2010).

The R MixSim package is used to obtain the simulated data (Maitra and Melnykov, 2010). Thirty simulated data sets are obtained in each of the 18 scenarios. Cluster analysis is performed using the Expectation-Maximization algorithm implemented in the Rmixmod package (Lebret et al, 2012) and the K-means algorithm implemented in the IBM SPSS Statistics software.

3 DATA ANALYSIS AND RESULTS

In this section we present the results referring to the simulated 3 clusters' data sets. The corresponding distributional parameters are presented in Tables 2 and 3. The results obtained refer to all scenarios previously indicated in section 2.

Table 2: Balanced simulated data sets distributional parameters.

Data set		Poor		Moderate		Weel	
Group	Variable	Mean	Var	Mean	Var	Mean	Var
1 (30%)	X1	10.5	3.5	11.9	1.1	10.5	1.0
	X2	2.3	0.5	2.5	0.3	2.5	1.3
	X3	7.8	2.0	8.0	0.9	4.3	1.8
2 (30%)	X1	10.0	3.0	9.8	1.2	15.0	2.2
	X2	2.5	0.3	1.5	0.3	4.0	1.2
	X3	7.0	1.0	6.8	0.7	7.0	1.5
3 (40%)	X1	9.5	2.0	11.8	1.4	7.0	2.3
	X2	2.0	0.4	2.0	0.4	6.2	1.6
	X_3	7.5	1.2	8.9	0.7	2.5	1.7
Average overlap		0.633		0.140		0.019	
Max. overlap		0.653		0.516		0.029	

Table 3: Unbalanced simulated data sets distributional parameters.

Data set		Poor		Moderate		Weel	
Group	Variable	Mean	Var	Mean	Var	Mean	Var
1 (60%)	X1	11.0	2.2	12.3	1.1	14.3	0.7
	X2	5.3	0.8	6.4	0.6	7.0	0.2
	X3	7.8	1.8	8.8	1.1	9.2	0.3
2 (30%)	X1	10.0	2.0	11.0	1.0	12.7	0.5
	X2	4.5	0.5	5.0	0.5	5.0	0.4
	X3	7.2	1.4	7.5	0.8	7.6	0.3
3 (10%)	X1	9.4	1.8	9.5	0.9	11.0	0.5
	X2	4.0	0.4	3.7	0.4	3.5	0.3
	X_3	7.0	1.5	6.6	0.7	6.0	0.2
Average overlap		0.632		0.143		0.021	
Max. overlap		0.868		0.215		0.115	

Table 4: IPA simulated, distributional and approximated expectations (values are averaged over the 30 datasets and correspond to external validation of EM clusters).

Dataset-	IPA	Dataset - separation								
		Poor			Moderate			Weel		
		sim	distrib	approx	sim	distrib	approx	sim	distrib	approx
Balanced	Rand	0.464	0.464		0.521	0.521		0.552	0.552	
	RR	0.209	0.209		0.148	0.148		0.115	0.115	
	GL	0.632		0.631	0.684		0.687	0.711		0.716
	J	0.275		0.270	0.232		0.224	0.204		0.195
	C	0.431	0.431		0.376	0.376		0.339	0.339	
	GK	0.000			0.000			0.000		
	SoS	0.212			0.228			0.224		
	SS2	0.160		0.140	0.132		0.106	0.114		0.086
	FM	0.453	0.453		0.381	0.381		0.339	0.339	
Unbalanced	Rand	0.500	0.500		0.505	0.505		0.504	0.504	
	RR	0.229	0.229		0.206	0.206		0.209	0.209	
	GL	0.666		0.668	0.671		0.673	0.670		0.672
	J	0.313		0.309	0.293		0.289	0.296		0.292
	C	0.476	0.476		0.453	0.453		0.457	0.457	
	GK	0.000			0.000			0.000		
	SoS	0.246			0.248			0.248		
	SS2	0.186		0.172	0.172		0.155	0.174		0.157
	FM	0.477	0.477		0.454	0.454		0.457	0.457	

Table 5: IPA simulated, distributional and approximated expectations (values are averaged over the 30 datasets and correspond to external validation of KM clusters).

Dataset-	IPA	Dataset – separation								
		Poor			Moderate			Weel		
		sim	distrib	approx	sim	distrib	approx	sim	distrib	approx
Balanced	R	0.552	0.552		0.551	0.551		0.552	0.552	
	RR	0.115	0.115		0.116	0.116		0.115	0.115	
	GL	0.711		0.716	0.710		0.715	0.711		0.716
	J	0.204		0.195	0.205		0.196	0.204		0.195
	C	0.339	0.339		0.341	0.341		0.339	0.339	
	GK	0.000			0.000			0.000		
	SoS	0.224			0.225			0.224		
	SS2	0.114		0.086	0.114		0.087	0.114		0.086
	FM	0.339	0.339		0.341	0.341		0.339	0.339	
Unbalanced	R	0.515	0.515		0.514	0.514		0.506	0.506	
	RR	0.152	0.152		0.158	0.158		0.198	0.198	
	GL	0.680		0.683	0.679		0.681	0.672		0.674
	J	0.239		0.231	0.246		0.238	0.285		0.280
	C	0.386	0.386		0.394	0.394		0.443	0.443	
	GK	0.000			0.000			0.000		
	SoS	0.235			0.237			0.246		
	SS2	0.136		0.110	0.140		0.115	0.166		0.148
	FM	0.390	0.390		0.398	0.398		0.444	0.444	

Table 6: IPA observed and adjusted Means and the corresponding coefficients of variation (values are averaged over the 30 datasets and correspond to external validation of EM clusters).

Dataset-	IPA	Dataset – separation											
		Poor				Moderate				Weel			
		obsM	cv	adjM	cv	obsM	cv	adjM	cv	obsM	cv	adjM	cv
Balanced	R	0.483	0.131	0.038	0.655	0.710	0.102	0.400	0.236	0.974	0.006	0.943	0.014
	RR	0.219	0.262	0.012	0.629	0.242	0.124	0.110	0.205	0.327	0.015	0.239	0.018
	GL	0.649	0.091	0.050	0.642	0.828	0.068	0.464	0.219	0.987	0.003	0.955	0.011
	J	0.293	0.108	0.024	0.658	0.459	0.085	0.293	0.249	0.927	0.018	0.908	0.022
	C	0.453	0.084	0.038	0.655	0.628	0.060	0.400	0.236	0.962	0.009	0.943	0.014
	GK	0.101	0.555	0.101	0.554	0.724	0.175	0.724	0.175	0.998	0.001	0.998	0.001
	SoS	0.234	0.199	0.028	0.636	0.485	0.146	0.333	0.257	0.943	0.014	0.927	0.018
	SS2	0.172	0.126	0.014	0.661	0.299	0.107	0.191	0.262	0.864	0.033	0.847	0.038
	FM	0.476	0.121	0.040	0.637	0.636	0.051	0.406	0.225	0.962	0.009	0.943	0.014
Unbalanced	R	0.605	0.062	0.211	0.362	0.847	0.025	0.690	0.064	0.990	0.004	0.980	0.009
	RR	0.282	0.152	0.069	0.373	0.377	0.056	0.215	0.072	0.452	0.029	0.307	0.021
	GL	0.753	0.039	0.260	0.348	0.917	0.014	0.747	0.053	0.995	0.002	0.985	0.006
	J	0.416	0.126	0.151	0.384	0.711	0.053	0.591	0.083	0.979	0.009	0.970	0.013
	C	0.585	0.091	0.211	0.362	0.830	0.032	0.690	0.064	0.989	0.005	0.980	0.009
	GK	0.409	0.338	0.409	0.338	0.934	0.025	0.934	0.025	1.000	0.000	1.000	0.000
	SoS	0.366	0.128	0.158	0.381	0.715	0.051	0.621	0.077	0.981	0.008	0.974	0.011
	SS2	0.264	0.156	0.096	0.403	0.553	0.078	0.460	0.107	0.959	0.018	0.950	0.021
	FM	0.587	0.093	0.212	0.362	0.831	0.032	0.690	0.063	0.989	0.005	0.980	0.09

Table 7: IPA observed and adjusted Means and the corresponding coefficients of variation (values are averaged over the 30 datasets and correspond to external validation of KM clusters).

Dataset-	IPA	Dataset – separation											
		Poor				Moderate				Weel			
		obsM	cv	adjM	cv	obsM	cv	adjM	cv	obsM	cv	adjM	cv
Balanced	R	0.567	0.007	0.035	0.272	0.704	0.019	0.341	0.083	0.968	0.008	0.929	0.017
	RR	0.123	0.024	0.009	0.273	0.193	0.037	0.087	0.080	0.323	0.014	0.235	0.018
	GL	0.724	0.005	0.044	0.269	0.826	0.011	0.400	0.075	0.984	0.004	0.944	0.014
	J	0.221	0.024	0.021	0.275	0.394	0.044	0.238	0.095	0.910	0.021	0.888	0.028
	C	0.362	0.020	0.035	0.272	0.565	0.032	0.341	0.083	0.953	0.011	0.929	0.017
	GK	0.077	0.269	0.077	0.268	0.635	0.062	0.635	0.062	0.997	0.001	0.997	0.001
	SoS	0.243	0.023	0.025	0.275	0.438	0.045	0.276	0.093	0.930	0.017	0.910	0.022
	SS2	0.124	0.027	0.012	0.278	0.246	0.055	0.148	0.106	0.836	0.039	0.815	0.045
	FM	0.362	0.020	0.035	0.272	0.565	0.032	0.341	0.082	0.953	0.011	0.929	0.017
Unbalanced	R	0.550	0.018	0.072	0.223	0.695	0.048	0.373	0.182	0.943	0.100	0.883	0.218
	RR	0.170	0.032	0.020	0.219	0.249	0.089	0.108	0.189	0.416	0.162	0.274	0.234
	GL	0.710	0.011	0.092	0.217	0.820	0.028	0.439	0.161	0.968	0.057	0.901	0.188
	J	0.274	0.029	0.046	0.228	0.450	0.108	0.272	0.219	0.889	0.196	0.850	0.270
	C	0.430	0.022	0.072	0.223	0.620	0.073	0.373	0.182	0.930	0.127	0.883	0.218
	GK	0.155	0.218	0.155	0.218	0.682	0.114	0.682	0.114	0.967	0.075	0.967	0.075
	SoS	0.274	0.032	0.051	0.231	0.469	0.105	0.304	0.207	0.895	0.184	0.862	0.250
	SS2	0.159	0.033	0.026	0.232	0.292	0.143	0.177	0.256	0.834	0.271	0.804	0.325
	FM	0.435	0.023	0.073	0.222	0.625	0.070	0.378	0.177	0.932	0.123	0.884	0.214

In Tables 4 and 5 we present the comparative precision of the proposed simulation based approach: the corresponding averages (under H_0) match the distributional averages whenever they are available - see (Albatineh, 2010) – and are similar to the approximated expected values - see (Albatineh and Niewiadomska-Bugaj, 2011). The correction of observed indices values, in Tables 6 and 7, obeys to formula (7).

The results regarding external validation of EM and KM clustering algorithms are reported in Tables 6 and 7. The diverse IPA are affected differently by the adjustment - the GL index is clearly the most affected by correction. Also, correction for change is particularly essential when considering poorly separated clusters.

As expected, the averages of simulated values, under H_0, of the GK index are null (Goodman and Kruskal, 1954). The R and C indices values are equal after adjustment which is in accordance with (Albatineh and Niewiadomska-Bugaj, 2006). We also conclude that, after adjustment, FM values are very similar to R and C values.

4 DISCUSSION AND PERSPECTIVES

In the present paper we focus on the correction of indices of paired agreement (IPA) between two partitions.

When comparing two partitions – e.g. when performing clustering validation and comparing clusters estimated and real clusters – agreement between them may be due to chance. This issue was first addressed by (Hubert and Arabie, 1985) referring to a specific measure of agreement - the Rand index of paired agreement. These authors provided a new adjusted Rand index excluding agreement by chance. Naturally, there are numerous IPA and this issue should be addressed when using any index. Recently, (Albatineh, 2010), for example, identified a family of paired indices and provided analytic formulas for their correction, using the corresponding averages under the hypothesis of independence. However, analytic correction cannot be provided for many indices – e.g. for the Jaccard index (a very old and well-known index) or the Gower and Legendre index, a more recent one.

As an alternative approach for IPA correction, we propose using the simulation of crosstabs to estimate the average of any index under the hypothesis of restricted independence i.e. subject to constraints of marginal totals (including the number of observations in the known clusters and the estimated ones). We generate 17,000 tables for the estimation of each average. Finally, we correct the observed IPA using their estimated average and use normalization so that all values can be compared. Nine IPA are analysed. The main contribution of this study is therefore to provide a method that is able to correct virtually any IPA for agreement by chance. When an analytic solution is available for correction (based on distributional assumptions), the differences between IPA analytic averages and averages provided by the proposed method are insignificant (at most 0.0001) which shows the method's precision.

To illustrate the usefulness of the proposed method for the indices' adjustment, we conduct external validation of the EM and KM algorithms within diverse scenarios.

According to the results obtained we identified notorious differences between the observed and adjusted indices when trying to capture a clustering structure originated in a poorly separated original mixture. This fact clearly demonstrates the pertinence of indices' correction. In fact, for difficult (impossible?) clustering tasks the observed indices clearly overestimate the clustering performance, while the adjusted indices translate the poor agreement with original clusters, despite of some variability which, we believe, is realistic.

For the moderately separated components, the agreement by chance factor yields minor correction to the paired indices, and when "easy" clusters (with a good separation) are considered, correction for chance is almost insignificant.

Performance of the EM algorithm is generally better. The gap between EM and KM is clearer in the case of unbalanced clusters. For "easy" clustering tasks, the KM and EM perform alike.

The results obtained underline the need to use adjusted indices, corrected for agreement by chance when conducting evaluation of (any) clustering algorithms' performance based on agreement with the original structure. Additional clustering algorithms and indices can be used in the future.

In future research, the distributions of alternative corrected indices should be further investigated for electing the most useful ones – those evidencing the least biased distributions and the easiest to interpret.

REFERENCES

Agresti, A., Wackerly, D. & Boyett, J. M., 1979. Exact conditional tests for cross-classifications: approximation of attained significance levels. *Psychometrika,* 44, 75-83.

Albatineh, A. N., Niewiadomska-Bugaj, M. & Mihalko, D., 2006. On Similarity Indices and Correction for Chance Agreement. *Journal of Classification,* 23, 301-313.

Albatineh, A. N., 2010. Means and variances for a family of similarity indices used in cluster analysis. *Journal of Statistical Planning and Inference,* 140, 2828-2838.

Albatineh, A. N. & Niewiadmska-Bugaj, M., 2011. Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification,* 5, 179-200.

Amorim, M. J. &Cardoso, M. G. M. S., 2010. Limiares De Concordância Entre Duas Partições. *Livro de Resumos do XVIII Congresso Anual da Sociedade Portuguesa de Estatística*, 47-49.

Amorim, M. J. P. C. & Cardoso, M. G. M. S., 2012. Clustering cross-validation and mutual information indices. *In: Ana Colubi, K. F., Gil Gonzalez-Rodriguesand Erricos John Kontoghiorghes, ed. 20th International Con-ference on Computational Statistics (COMPSTAT 2012), 2012 Limassol, Cyprus. The International Statistical Institute/International Association for Statistical Computing*, 39-52.

Chumwatana, T., Wong, K. W. & Xie, H., 2010. A SOM-Based Document Clustering Using Frequent Max Substrings for Non-Segmented Texts. *J. Intelligent Learning Systems & Applications,*, 2, 117-125.

Czekanowski, J., 1932. "Coefficient of racial likeness" and "durchschnittliche Differenz". *Anthropologischer Anzeiger,* 14, 227-249.

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm.*Journal of the Royal Statistical Society.* Series B (Methodological), 1-38.

Everit, B., Landau, S. & Leese, M. 2001. *Cluster Analysis,* London, Arnold.

Fowlkes, E. B. &mallows, C. L., 1983. A method for comparing two hierarchical clusterings.*Journal of the American Statistical Association*, 78, 553-569.

Goodman, L. A. & Kruskal, W. H., 1954. Measures of Association for Cross Classifications. *Journal of the American Statistical Associations,* 49.

Gower, J. C. & Legendre, P., 1986. Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification,* 3.

Halton, J. H., 1969. A rigorous derivation of the exact contingency formula. *In:Proceedings of the Cambridge Philosophical Society*. Cambridge Univ Press, 527-530.

Hennig, C., 2006. Cluster-wise assessment of cluster stability. *Research report n° 271,* Department of Statistical Science, University College London.

Hubert, L. and Arabie, P. 1985. Comparing partitions. *Journal of classification,* 2, 193-218.

Jaccard, 1908. Nouvelles Recerches sur la Distribuition Florale. *Bulletin de la Societé Vaudoise de Sciences Naturells,* 44, 223-370.

Jain, A. K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters,* 31, 651-666.

Krzanowski, W. J. & Marriott, F. H. C., 1994. *Multivariate analysis*, Edward Arnold London.

Lebret, R., S., L., Langrognet, F., Biernacki, C., Celeux, G. & Govaert, G., 2012. Rmixmod: The r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library.http://cran.r-project.org/web/ packages/Rmixmod/index.html.

Maitra, R. & Melnykov, V., 2010. Simulating data to study performance of finite mixture modeling and clustering algorithms. *Computational and Graphical Statistics,* 19, 354-376.

Meyeri, A. D. S., Garcia, A. A. F., Souza, A. P. & JR., C. L. D. S., 2005. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L). *Genetics and Molecular Biology,* 27, 83-91.

Milligan, G. W. & Cooper, M. C., 1986. A Study of Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Reserch,* 21, 441-458.

O'Hagan, A., Murphy, T. B. & Gormley, I. C., 2012. Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Computational Statistics and Data Analysis,* 56, 3843-3864.

Rand, W. M., 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association,* 66, 846-850.

RusseL, P. F. & Rao, T. R. 1940. On Habitat and Association of Species of Anophelinae Larvae in South-Eastern Madras. *J. Malar. Inst. India,* 3, 153-178.

Shamir, O. and tishby, N., 2010. Stability and model selection in k-means clustering. *Mach Learn,* 80, 213-244.

Sokal, R. R. and Sneath, P. H., 1963. *Principles of Numerical Taxonomy,* San Francisco CA: Freeman.

CHAPTER 4: COMPARING CLUSTERING SOLUTIONS: THE USE OF ADJUSTED PAIRED INDICES

This manuscript has the following reference:

Amorim, M. J. & Cardoso, M. G. M. S. 2015c. Comparing clustering solutions: The use of adjusted paired indices. Intelligent Data Analysis, 19, 1275–1296.

Intelligent Data Analysis 19 (2015) 1275–1296
DOI 10.3233/IDA-150782
IOS Press

Comparing clustering solutions: The use of adjusted paired indices

Maria José Amorim[a] and Margarida G.M.S. Cardoso[b,*]
[a]*Department of Mathematics of ISEL and UNIDE Lisboa, Portugal*
[b]*UNIDE and Department of Quantitative Methods for Management and Economics of ISCTE, Lisbon University Institute, Lisbon, Portugal*

Abstract. In the present paper we compare clustering solutions using indices of paired agreement. We propose a new method – IADJUST – to correct indices of paired agreement, excluding agreement by chance. This new method overcomes previous limitations known in the literature as it permits the correction of any index. We illustrate its use in external clustering validation, to measure the accordance between clusters and an *a priori* known structure. The adjusted indices are intended to provide a realistic measure of clustering performance that excludes agreement by chance with ground truth. We use simulated data sets, under a range of scenarios – considering diverse numbers of clusters, clusters overlaps and balances – to discuss the pertinence and the precision of our proposal. Precision is established based on comparisons with the analytical approach for correction – specific indices that can be corrected in this way are used for this purpose. The pertinence of the proposed correction is discussed when making a detailed comparison between the performance of two classical clustering approaches, namely Expectation-Maximization (EM) and K-Means (KM) algorithms. Eight indices of paired agreement are studied and new corrected indices are obtained.

Keywords: Adjusted indices, indices of paired agreement, clustering evaluation, external evaluation

1. Motivation and objectives

In the present study we focus on the use of indices of paired agreement to measure the accordance between two partitions of the same data set: we propose a new method to correct the indices that handles agreement by chance.

More specifically we consider eight indices of paired agreement to illustrate the precision and pertinence of our approach: Jaccard [19], Czekanowski [10], Goodman and Kruskal [13], Sokal and Sneath [33] (2 indices), Rand [32], Fowlkes and Mallows [12] and Gower and Legendre [14]. All indices considered refer to paired agreement – e.g. the Rand index quantifies the proportion of pairs of observations both partitions agree to put in the same cluster or separate in different clusters – and their maximum value is one.

The problem can be illustrated by the following example. Consider the data set Statlog (Vehicle Silhouettes) in the UCI Machine Learning Repository [5]. Built in 1986/87, the data set refers to images that a camera at a specific angle (silhouettes) obtains of four types of vehicles (classes): a double decker

[*] Corresponding author: Margarida G.M.S. Cardoso, UNIDE and Department of Quantitative Methods for Management and Economics of ISCTE – Lisbon University Institute, Lisboa 1649-026, Portugal. Tel.: +351 21 7903264; E-mail: margarida.cardoso@iscte.pt.

bus, a Chevrolet van, a Saab 9000 and an Opel Manta 400. Statlog attributes refer to diverse image characteristics of the bus, van and two cars.

In an attempt to (externally) evaluate whether an Expectation-Maximization (EM) based clustering algorithm [23] is able to recover the data true classes using the image attributes available, we run this algorithm and obtain the confusion matrix in Table 1.

Although a naïve evaluation of the performance of EM (based on Table 1) is provided by the percentage of observations that are correctly allocated to the original classes by the clustering algorithm (44.9%), this is a poor measure of performance – e.g. [34]. Examples of alternative measures of (paired) agreement between true classes and uncovered groups, based on all the confusion matrix cells include the Gower and Legendre index (*GL*), the Sokal and Sneath index (*SoS*) and the well-known Rand index (R). In the Statlog case, $GL = 0.804$, $SoS = 0.283$ and $R = 0.673$. So, what can we say about the proportions of pairs of observations that both partitions agree to join and/or to separate? Is it high (the *GL* index is near one) or low (the SoS index is near zero) or moderate (according to Rand index)? Are there specific thresholds we should consider for each index?

Hubert and Arabie [18] addressed this threshold problem for the Rand index and proposed a new adjusted Rand (R_a) index which excludes agreement by chance. The generic adjusted index is:

$$R_a = \frac{R - E_0\left(R\right)}{1 - E_0\left(R\right)} \tag{1}$$

where $E_0\left(R\right)$ stands for the expected value of the Rand index under the null hypothesis (H_0) of agreement by chance between the partitions. This adjusted Rand is the ratio between the observed improvement and the maximum possible improvement, taking as a reference the average value of the index under H_0. Thus, adjusted Rand is expected to have a null value when all agreement between two partitions is due to chance and its maximum value is one.

When applied to our example (i.e. for the Statlog results in Table 1), the Hubert and Arabie adjusted Rand index is $R_a = 0.143$. This value is clearly distant from the uncorrected one ($R = 0.673$), referred to above, and is now easy to interpret: it represents a weak EM algorithm recovery of the true Statlog classes that is beyond chance level.

Milligan and Cooper [29] rapidly valued the R_a index for its desirable properties, namely in the context of clustering external validation. Meanwhile, there have been countless proposals to obtain corrected versions of several indices. However, as we will see in the literature review on the correction of indices for agreement by chance, there are still indices with no corrected version available. This is because the analytical approach – e.g. [37] and [2] is unable to deal with all the indices. Moreover, despite some attempts to use approximation approaches, [3], these too are not available for all indices.

Hence, in the Statlog example, index *GL* could only be corrected using an approximation approach and, to our knowledge, no method is available to correct the *SoS* index.

The present contribution aims to fill this gap in the literature by proposing a simulation based method – IADJUST – which is able to adjust any index of paired agreement, excluding agreement by chance. The R software is used for its implementation.

We conduct extensive numerical experiments to assert the precision and pertinence of IADJUST. We then revisit the Statlog data set example and compare all the IADJUST corrected indices values, as opposed to the values of the plain indices, to determine whether the EM algorithm used can uncover the Statlog four classes.

Finally, we discuss new insights concerning the adequate correction of paired indices in light of the experimental results obtained.

Table 1
Confusion matrix between EM clustering results and Statlog classes

	C1	C2	C3	C4	Total
Bus	*127*	42	48	1	218
Opel	31	*103*	44	34	212
Saab	30	99	*53*	35	217
Van	91	0	11	*97*	199
Total	279	244	156	167	846

Table 2
Similarity matrix (counts of pairs of observations)

Partition P^K	Partition P^Q	
	Pair in same cluster	Pair not in same cluster
Pair in same cluster	a_{11}	a_{10}
Pair not in same cluster	a_{01}	a_{00}

Table 3
Indices of paired agreement under study

IPA		$\mathscr{L}$ Family	Formula	Reference	Correction
R	Rand	√	$\frac{a_{11}+a_{00}}{a_{11}+a_{10}+a_{01}+a_{00}}$	[32]	DIST
GL	Gower and Legendre	×	$\frac{a_{11}+a_{00}}{a_{11}+0.5\,(a_{10}+a_{01})+a_{00}}$	[14]	APPROX
J	Jaccard	×	$\frac{a_{11}}{a_{11}+a_{10}+a_{01}}$	[19]	APPROX
C	Czekanowski	√	$\frac{2a_{11}}{2a_{11}+a_{10}+a_{01}}$	[10]	DIST
GK	Goodman and Kruskal	×	$\frac{a_{11}a_{00}-a_{10}a_{01}}{a_{11}a_{00}+a_{10}a_{01}}$	[13]	(no need)
SoS	Sokal and Sneath	×	$\frac{a_{11}a_{00}}{\sqrt{(a_{11}+a_{10})\,(a_{11}+a_{01})\,(a_{00}+a_{10})\,(a_{00}+a_{01})}}$	[33]	IADJUST
SS2	Sokal and Sneath (2)	×	$\frac{a_{11}}{a_{11}+2\,(a_{10}+a_{01})}$	[33]	APPROX
FM	Fowlkes and Mallows	√	$\frac{a_{11}}{\sqrt{(a_{11}+a_{10})\,(a_{11}+a_{01})}}$	[12]	DIST

2. Using indices of paired agreement between two partitions

2.1. Common indices used in clustering validation

Several similarity indices can be used to measure the agreement between two partitions of the same data – P^K and P^Q with K and Q groups, respectively – generally designated Indices of Agreement (IA) – e.g. [14,29]. These indices are used in various domains – e.g. see [20].

In this work we address specific IA that are based on the number of pairs of observations that both partitions allocate (or not) to the same cluster – Indices of Paired Agreement (IPA). IPA can be determined from a similarity matrix (Table 2). They can be seen as measures of dissimilarity between two binary variables, indicating whether or not a pair of observations is in the same cluster (situations that are typically coded 1 and 0, respectively). Therefore, many other IPA might be considered – e.g. see [39].

The specific IPA under study are in Table 3 – although limited in number, they allow us to illustrate the IPA diversity. All the indices in Table 3 can also be written based on the elements of the confusion matrix between partitions P^K and P^Q being compared: a $K \times Q$ matrix, whose $(k,q)^{\text{th}}$ element, n_{kq}, is the number of observations in the intersection of group C^k of $P^K(k = 1 \ldots K)$ and C^q of $P^Q(q = 1 \ldots Q)$; n_{k+} and n_{+q} represent the rows and columns totals (respectively) and n the number of observations. The values of a_{11}, a_{10}, a_{01} and a_{00} in Eqs (2)–(5) establish the link between the IPA and the confusion matrix cells:

$$a_{11} = \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^2 - \frac{n}{2} \tag{2}$$

$$a_{10} = \frac{1}{2}\sum_{k=1}^{K} n_{k+}^2 - \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^2 \tag{3}$$

$$a_{01} = \frac{1}{2}\sum_{q=1}^{Q} n_{+q}^2 - \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^2 \tag{4}$$

$$a_{00} = \binom{n}{2} - a_{11} - a_{10} - a_{01} \tag{5}$$

Thus, for example, the Rand index formula (see Table 3), can be rewritten as:

$$R\left(P^K, P^Q\right) = \frac{\binom{n}{2} + \sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2 - \frac{1}{2}\left(\sum_{q=1}^{Q} n_{+q}^2 + \sum_{k=1}^{K} n_{k+}^2\right)}{\binom{n}{2}} \tag{6}$$

In the context of clustering validation, IPA are used to measure the agreement between partitions drawn from slightly modified data sets (to decide upon a clustering solution stability), or to measure the agreement between clustering solutions and the real partition (external validation). In external validation, the clustering solutions are ranked according to their degree of agreement with ground truth. The following references illustrate the use of IPA for this purpose.

Melnylkov and Melnylkov [28] proposed a strategy for initializing the EM algorithm, when estimating multivariate Gaussian mixture models, and used the Hubert and Arabie adjusted Rand index, R_a to evaluate its comparative performance with diverse initializations of the EM algorithm. They used simulated data sets of size 10,000 drawn from a two-dimensional Gaussian mixture with 20 components, and a real data set (a color image quantization problem) to conduct clustering external validation.

McNicholas and Murphy [26], also used the R_a, index. These authors compared the performance of several clustering procedures (including diverse Gaussian mixture models and the K-means algorithm) applied to two real gene expression data sets.

In Lu et al. [24], the R_a, the Jaccard and the Wallace IPA [35], were used as objective functions of clustering ensemble. These authors used the R_a instead of the R index since "A problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value" (p. 667).

In general, the Hubert and Arabie adjusted Rand index is increasingly used in the context of clustering validation. The success of this index is primarily due to its capacity to exclude agreement by chance when quantifying (paired) agreement between two partitions. However, several adjusted IPA – as opposed to plain IPA (like those in Table 3) – are now available and therefore diverse competitors of the R_a index.

2.2. *Correcting paired indices for agreement by chance*

2.2.1. *The kappa index*

When comparing two partitions, their agreement, which is quantified based on the corresponding confusion matrix, can be due to chance. Therefore, in order to adequately evaluate the degree of agreement between two partitions, IA must be corrected and their expected value under agreement by chance (H_0) should be excluded (revisit Eq. (1)).

The issue of agreement by chance between two nominal variables was first raised by Cohen [9] interpreting the index of percent agreement. This author suggested adjusting the index of percentage agreement and obtained the *kappa* index which quantifies the distancing from a situation of independence (H_0) and goes to zero when the agreement occurs by chance alone. In fact, to obtain the *kappa* index, Eq. (1) is applied to percentage agreement:

$$\textit{kappa}\left(P_a^K, P_b^K\right) = \textit{Perc}_a\left(P_a^K, P_b^K\right) = \frac{\sum_{k=1}^{K} \frac{n_{kk}}{n} - \sum_{k=1}^{K} \frac{n_{k+}n_{+k}}{n^2}}{1 - \sum_{k=1}^{K} \frac{n_{k+}n_{+k}}{n^2}} \tag{7}$$

Note that in the *kappa* index calculus it is assumed there is specific correspondence between the clusters of both partitions and also a common number of clusters (K).

2.2.2. The adjusted rand index

Following the adjustment of the rate of percentage agreement, other IA were corrected using the hypothesis of independence between the two partitions being compared. In the context of paired agreement, Brennan and Light [7] were the first to address this issue. Hubert and Arabie [18], explored the same idea of paired agreement to propose an adjusted version of the Rand index. For correction, they considered the mean of this index under the null hypothesis (H_{0ic}) of no association between the partitions being compared, conditional on the row and column table totals, namely the hypothesis of restricted independence. Under H_{0ic} the probability of observing a specific confusion matrix can be modeled by the Generalized Hypergeometric distribution [15]. Each matrix cell follow an Hypergeometric distribution and thus $E_{0ic}\left[R\left(P^K, P^Q\right)\right]$ can be calculated as follows (see also Eq. (6) and note that $\sum_{q=1}^{Q} n_{+q}^2$ and $\sum_{k=1}^{K} n_{k+}^2$ as well as $\binom{n}{2}$ are constants):

$$E_{0ic}\left[R\left(P^K, P^Q\right)\right] = \frac{\binom{n}{2} + E_0\left[\sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2\right] - \frac{1}{2}\left(\sum_{q=1}^{Q} n_{+q}^2 + \sum_{k=1}^{K} n_{k+}^2\right)}{\binom{n}{2}} \tag{8}$$

where

$$E_{0ic}\left[\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^2\right] = \frac{\sum_{k=1}^{K}\sum_{q=1}^{Q} n_{k+}^2 n_{+q}^2}{n\left(n-1\right)} - \frac{\sum_{k=1}^{K} n_{k+}^2 + \sum_{q=1}^{Q} n_{+q}^2}{n-1} + \frac{n^2}{n-1} \tag{9}$$

The Hubert and Arabie adjusted Rand index, R_a, is then obtained combining Eqs (1), (8) and (9). It is bounded by 1 and takes the value zero when the observed index R is equal to its expected value under H_{0ic}.

The merit of R_a was first highlighted in a study by Milligan and Cooper [29] on the external validation of results of hierarchical clustering: the R_a was found to outperform four other indices. When Milligan and Cooper considered the Rand index [32], the Jaccard [19], the Fowlkes and Mallows [12] and another adjusted version of the Rand index [30], they concluded after extensive numerical experiments, that R_a was the best external criterion for clustering evaluation since its values for the simulated scenarios of agreement by chance were consistently close to zero.

2.2.3. *Adjusted paired indices: The analytical approach*

The Hubert and Arabie adjusted Rand Eq. (1) can be extended to any index of paired agreement – replacing R by any *IPA* in the formula; unlike the *IPA* values, the values of the different IPA_a can be compared, since a similar scale is used. In order to correct any IPA $E_0\left(IPA\right)$ must therefore be determined.

Theoretically, under H_{0ic}, any exact IPA mean can be determined considering all possible confusion matrices under the hypothesis of restricted independence. However, this is only feasible for relatively small tables with small numbers of observations, due to computational complexity [22].

In order to obtain an IPA mean we can also resort to a distributional model, as in Hubert and Arabie [18]. Using this approach specifically for the Goodman and Kruskal index, the expected value under H_{0ic} is null $E_{0ic}\left[GK\left(P^K, P^Q\right)\right] = 0$ [8,37,40]. The proof is easily obtained using the equivalence between statistical independence in accordance with the generalized hypergeometric distribution, used in [18] for the contingency table between two partitions, and the statistical independence according to the binomial distribution, for the pairwise concordance table (i.e. Table 2) – [36]. As such, the counts in the GK formula (in Table 3) simply have to be replaced by their corresponding expected values under H_{0ic} – e.g. replace a_{11} by:

$$E_{0ic}\left[a_{11}\right] = \frac{\left(a_{11}+a_{10}\right)\left(a_{11}+a_{01}\right)}{a_{11}+a_{10}+a_{01}+a_{00}} \tag{10}$$

It is worth noting that the distribution of the n_{kq} depends upon whether the marginal totals of the confusion table are assumed to be fixed (Hypergeometric model) or variable (multinomial model). In the present study we take the marginals to be fixed (following [18]): thus, henceforth, H_0 is H_{0ic}. Nevertheless, as [7] pointed out (citing [21]), it should also be noted that "for large samples both assumptions lead to the same large-sample distribution for the cell entries" (p. 157).

Families of indices may also be used to provide the calculation of the respective expected values. Albatineh et al. [4] presented a family of IPA, the $\mathscr{L}$ Family, that are linear functions of the sum of the squares of the n_{kq} cells of the confusion matrix:

$$IPA = \alpha + \beta \sum_{k=1}^{K}\sum_{q=1}^{Q} n_{kq}^2 \tag{11}$$

where α and β depend on the marginal counts and are unique for each index.

The Rand, Czekanowski and Fowlkes and Mallows indices, under study, belong to this family and therefore their means and variances may be viewed as special cases of a general formula based on the n_{kq} cells distribution (see DIST in "Correction" column in Table 3). For the Rand index, for example:

$$\alpha = 1 - \frac{1}{n\left(n-1\right)}\left(ssq_{k+} + ssq_{+q}\right) \tag{12}$$

$$\beta = \frac{2}{n\left(n-1\right)} \tag{13}$$

where

$$ssq_{k+} = \sum_{k=1}^{K} n_{k+}^2 \tag{14}$$

and

$$ssq_{+q} = \sum_{q=1}^{Q} n_{+q}^2 \tag{15}$$

The expected values of the $\mathscr{L}$ family indices all obey a general formula [4]. Using this framework it can also be proved that the Rand and the Czekanowski indices become equivalent after correction by chance since they have the same $\frac{1-\alpha}{\beta}$ ratio [4]. In fact, the Rand and the Czekanowski indices become the S_{Cohen} index [9] after correction for chance [36].

Warrens [37–39] expressed the $\mathscr{L}$ family differently, i.e. as a linear function of the observed proportions of agreement corresponding to the cell entries in Table 2:

$$IPA = \lambda + \mu\,(p_{11} + p_{00}) \tag{16}$$

In this Eq. (16), p_{11} is the proportion of pairs that both partitions agree to put in the same cluster – see Eq. (17) – p_{00} is the proportion of pairs that both partitions agree to allocate in different clusters and λ and μ depend on the marginal distributions and are defined for each $\mathscr{L}$ family index.

$$p_{11} = \frac{a_{11}}{a_{11} + a_{10} + a_{01} + a_{00}} \tag{17}$$

Since linearity in $(p_{11} + p_{00})$ is equivalent to linearity in p_{11} or in p_{00}, Eq. (16) can be rewritten:

$$IPA = \kappa + \nu\,(p_{11}) \tag{18}$$

where κ and ν depend on the marginal distributions and are different for each index, [37].

2.2.4. Approximation methods for correcting paired indices

The distribution based analytical solutions to determine $E_0\,(IPA)$, are only available for a few indices. However, taking advantage of the relationships between various indices, the calculus of the approximated expected values, under H_0, can be attained for a few more indices. For the Jaccard, Gower and Legendre and Sokal and Sneath (SS2) indices, Albatineh et al. [3] conducted regressions, based on simulated data (samples of 500 observations with five balanced clusters based on two Gaussians were considered), and estimated that:

$$\hat{J} = 0.656151C^2 + 0.298540C + 0.018352 \tag{19}$$

$$\widehat{GL} = -0.656151R^2 + 1.610843R + 0.026957 \tag{20}$$

$$\widehat{SS2} = 1.10870C^2 - 0.28021C + 0.05337 \tag{21}$$

Equations (19)–(21) allow the correction of three indices under study to be approximated (see APPROX in column "Correction" in Table 3). For the Jaccard index, for example, the corresponding expected value is derived based on Eq. (19) (note that the expected value of the Czekanowski index are easily derived as it belongs to the $\mathscr{L}$ family). Previously, [34] had also conducted simple linear regressions that established the relationships between indices of paired agreement, namely the Adjusted Rand index (viewed as a landmark in the study of external assessment of clustering results) and the Rand, Jaccard and Fowlkes and Mallows indices – however, and despite conducting an extensive experimental study, no attempt was made to correct any other index.

An alternative approximation approach proposed in [3] resorts to Taylor series. However, according to the same authors [3], its accuracy is lower than the one obtained via regression.

2.2.5. A simulation approach for indices' correction

Despite the diverse approaches to handle the correction for agreement by chance, a number of IPA are not covered by the procedures so far proposed, namely the Sokal and Sneath (*SoS*) index under study [4]. To overcome this limitation we propose a new methodology – IADJUST – that can deal with the correction of any IPA for agreement by chance.

In order to illustrate the precision of the proposed approach we provide comparisons with the analytical-based approach for correction of the Rand [32], Czekanowski [10] and Fowlkes and Mallows [12] indices. Comparisons are also provided with the approximation approach for the Jaccard index [19], one of the Sokal and Sneath [33] indices and the Gower and Legendre [14] index.

The usefulness of the proposed method is illustrated when evaluating the performance of two well-known clustering tools: the Expectation Maximization (EM) [11], implemented in the Rmixmod package [23], and the Hartigan K-Means (KM) algorithm [16], within diverse experimental scenarios.

The numerical experiments conducted resort to simulated data sets and consider the following experimental factors: the number of clusters (with corresponding fixed dimension), the degree of clusters' overlap and balance.

3. The proposed iadjust method

3.1. IADJUST step by step

In the present work, we propose a new method – IADJUST – to correct all indices of paired agreement. The method provides the average value of any IPA under the null hypothesis of agreement due to chance, for use in its correction (as in Eq. (1) applied to the Rand index). We therefore obtain new adjusted Jaccard, Gower and Legendre and Sokal and Sneath indices and provide means to correct any other index of agreement.

IADJUST also provides the indices' empirical distribution under the null hypothesis. Therefore, we expect to glean new insights regarding the comparative performance of the indices under study, specifically within the experimental numerical settings considered (see next section).

Under the hypothesis of restricted independence – H_{0ic}, simply designated H_0 – the probability of observing a specific confusion matrix can be modeled by the Generalized Hypergeometric distribution [15] and the variables $(N_{k1}, N_{k2}, \ldots, N_{kq})$ (counts associated to each line of the confusion matrix) are independent multinomial random variables ($k = 1, 2, \ldots, K$) with $n_{k+} = \sum_{q=1}^{Q} n_{kq}$ and all with the same probability vector parameters. Given the values in the previous rows and columns, the conditional probability of the confusion matrix $(lm)^{\text{th}}$ cell value $f\left(n_{lm} \mid n_{+m}; n_{l+}\right)$ is given by the Hypergeometric distribution, e.g. [31], with parameters: $\sum_{q=m}^{Q}\left(n_{+q} - \sum_{k=1}^{l-1} n_{kq}\right)$ (population size) $n_{l+} - \sum_{q=1}^{m-1} n_{+q}$ (number of successes in the population) and $n_{+m} - \sum_{k=1}^{l-1} n_{km}$ (sample size) i.e.

$$N_{lm} \sim H\left(\sum_{q=m}^{Q}\left(n_{+q} - \sum_{k=1}^{l-1} n_{kq}\right); n_{l+} - \sum_{q=1}^{m-1} n_{+q}; n_{+m} - \sum_{k=1}^{l-1} n_{km}\right) \tag{22}$$

$$f\left(n_{lm} \mid n_{+m}; n_{l+}\right) \tag{23}$$

$$= \frac{\left(n_{l+} - \sum_{q=1}^{m-1} n_{lq}\right)! \left(n - \sum_{k=1}^{l} n_{k+} - \sum_{q=1}^{m-1} n_{+q} + \sum_{q=1}^{m-1} \sum_{k=1}^{l} n_{kq}\right)!}{n_{lm}! \left(n_{+m} - \sum_{k=1}^{l} n_{km}\right)! \left(n_{l+} - \sum_{q=1}^{m} n_{lq}\right)!}$$

$$\times \frac{\left(n_{+m}-\sum_{k=1}^{l-1} n_{km}\right)!\left(\sum_{q=m+1}^{Q}\left(n_{+q}-\sum_{k=1}^{l-1} n_{kq}\right)\right)!}{\left(n-\sum_{k=1}^{l} n_{k+}-\sum_{q=1}^{m} n_{+q}+\sum_{q=1}^{m}\sum_{k=1}^{l} n_{kq}\right)!\left(\sum_{q=m}^{Q}\left(n_{+q}-\sum_{k=1}^{l-1} n_{kq}\right)\right)!}$$

Where $1 \leqslant l \leqslant K-1$ and $1 \leqslant m \leqslant Q-1$. If $l = K$ or $m = Q$ then n_{lm} can take only one possible value (its probability being one).

The (conditional) expected value of N_{lm} given previous entries (n_{kq}; $k = 1, \ldots, l-1$; $q = 1, \ldots, m-1$) and the row and column totals is:

$$E\left[N_{lm} \mid n_{+m}; n_{l+}\right] = \frac{\left(n_{+m}-\sum_{k=1}^{l-1} n_{km}\right)\left(n_{l+}-\sum_{q=1}^{m-1} n_{lq}\right)}{\sum_{q=m}^{Q}\left(n_{+q}-\sum_{k=1}^{l-1} n_{kq}\right)} \tag{24}$$

If the denominator in Eq. (24) is zero then $E\left[n_{lm} \mid n_{+m}; n_{l+}\right] = 0$.

The IADJUST procedure capitalizes on the generation of multiple confusion matrices, under H_0. The procedure starts from a confusion matrix that summarizes the agreement between two partitions and proceeds as illustrated in Table 4. In the present study, the IADJUST input applies to the comparison of clusters vs. classes known *a priori*, providing the external validation of clustering algorithms: EM and KM.

In order to implement the generation of each confusion matrix, under H_0, we resort to a new implementation of the Patfield procedure [31], using the R software. The proposed procedure is summarized in Table 5. An alternative procedure was previously reported in [34]. It resorted to an arbitrary scheme to generate the confusion matrix which was initialized with a diagonal matrix in which the diagonal cells concentrate all observations of *a priori* classes. These cells were progressively scattered into the off-diagonal cells and one hundred confusion matrices were generated for each experimental scenario considered. This study thus required a fixed partition (*a priori* known classes) which suits the external validation of a partition but not the evaluation of its stability. Moreover, it did not consider a particular distributional model for H_0. The proposal of IADJUST generation of confusion matrices obeys a specific statistical distributional model (Generalized Hypergeometric distribution).

Based on the generated confusion matrices, we obtain the average values of indices under H_0. In order to establish precision, we use the following rational. Let G be the number of generated matrices with corresponding "generated IPA value greater than the observed IPA" ("success"). G has a Binomial distribution with T trials and "success" probability π. If T is large, then G and G/T (the estimated p-value) are approximately Normally distributed, with G having mean $T\pi$ and variance $T\pi(1-\pi)$, and G/T having mean equal to π and variance $\pi(1-\pi)/T$. Therefore, Eq. (25) presents a $100(1-\alpha)\%$ confidence interval for π (where $\Phi^{-1}(1-\alpha/2)$ denotes the $(1-\alpha/2)$ percentile of the standard normal distribution):

$$\left(\hat{\pi}-\Phi^{-1}(1-\alpha/2)\sqrt{\frac{\pi(1-\pi)}{T}}, \hat{\pi}+\Phi^{-1}(1-\alpha/2)\sqrt{\frac{\pi(1-\pi)}{T}}\right) \tag{25}$$

If we want to estimate π with $100(1-\alpha)\%$ confidence and a maximum error equal to B, and since $\pi(1-\pi) \leqslant \frac{1}{4}$, we can resort to Eq. (26):

$$T \geqslant \frac{1}{4}\left[\frac{\Phi^{-1}(1-\alpha/2)}{B}\right]^2 \tag{26}$$

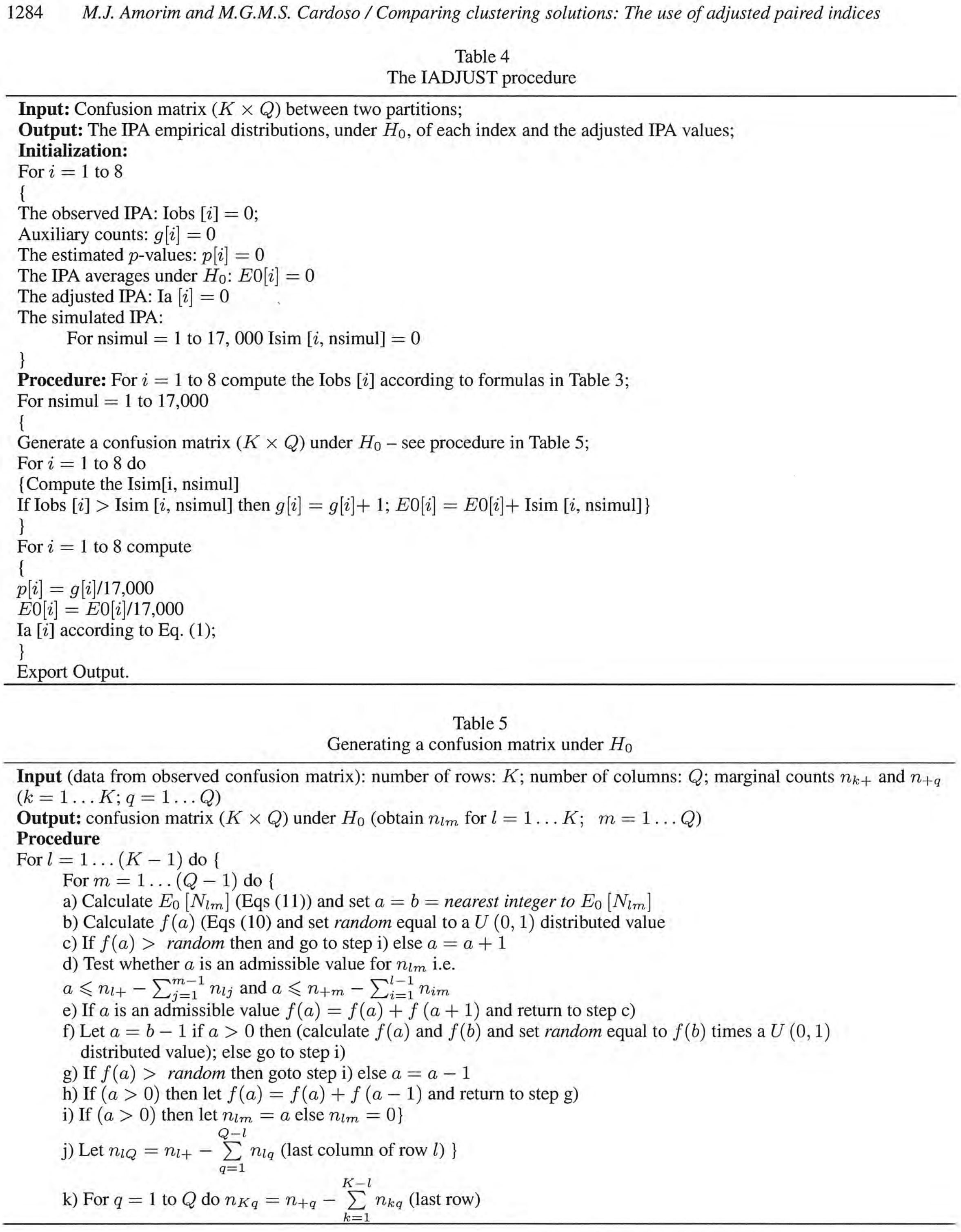

Table 4
The IADJUST procedure

```
**Input:** Confusion matrix ($K \times Q$) between two partitions;
**Output:** The IPA empirical distributions, under $H_0$, of each index and the adjusted IPA values;
**Initialization:**
For $i = 1$ to 8
{
The observed IPA: Iobs [$i$] $= 0$;
Auxiliary counts: $g[i] = 0$
The estimated $p$-values: $p[i] = 0$
The IPA averages under $H_0$: $E0[i] = 0$
The adjusted IPA: Ia [$i$] $= 0$
The simulated IPA:
        For nsimul $=$ 1 to $17,000$ Isim [$i$, nsimul] $= 0$
}
**Procedure:** For $i = 1$ to 8 compute the Iobs [$i$] according to formulas in Table 3;
For nsimul $=$ 1 to 17,000
{
Generate a confusion matrix ($K \times Q$) under $H_0$ – see procedure in Table 5;
For $i = 1$ to 8 do
{Compute the Isim[i, nsimul]
If Iobs [$i$] $>$ Isim [$i$, nsimul] then $g[i] = g[i]+ 1$; $E0[i] = E0[i]+$ Isim [$i$, nsimul]}
}
For $i = 1$ to 8 compute
{
$p[i] = g[i]$/17,000
$E0[i] = E0[i]$/17,000
Ia [$i$] according to Eq. (1);
}
Export Output.
```

Table 5
Generating a confusion matrix under H_0

```
**Input** (data from observed confusion matrix): number of rows: $K$; number of columns: $Q$; marginal counts $n_{k+}$ and $n_{+q}$
($k = 1 \ldots K; q = 1 \ldots Q$)
**Output:** confusion matrix ($K \times Q$) under $H_0$ (obtain $n_{lm}$ for $l = 1 \ldots K; \quad m = 1 \ldots Q$)
**Procedure**
For $l = 1 \ldots (K - 1)$ do {
    For $m = 1 \ldots (Q - 1)$ do {
    a) Calculate $E_0 [N_{lm}]$ (Eqs (11)) and set $a = b =$ *nearest integer to* $E_0 [N_{lm}]$
    b) Calculate $f(a)$ (Eqs (10) and set *random* equal to a $U(0, 1)$ distributed value
    c) If $f(a) >$ *random* then and go to step i) else $a = a + 1$
    d) Test whether $a$ is an admissible value for $n_{lm}$ i.e.
    $a \leqslant n_{l+} - \sum_{j=1}^{m-1} n_{lj}$ and $a \leqslant n_{+m} - \sum_{i=1}^{l-1} n_{im}$
    e) If $a$ is an admissible value $f(a) = f(a) + f(a + 1)$ and return to step c)
    f) Let $a = b - 1$ if $a > 0$ then (calculate $f(a)$ and $f(b)$ and set *random* equal to $f(b)$ times a $U(0, 1)$
       distributed value); else go to step i)
    g) If $f(a) >$ *random* then goto step i) else $a = a - 1$
    h) If $(a > 0)$ then let $f(a) = f(a) + f(a - 1)$ and return to step g)
    i) If $(a > 0)$ then let $n_{lm} = a$ else $n_{lm} = 0$}
    j) Let $n_{lQ} = n_{l+} - \sum_{q=1}^{Q-l} n_{lq}$ (last column of row $l$) }
    k) For $q = 1$ to $Q$ do $n_{Kq} = n_{+q} - \sum_{k=1}^{K-l} n_{kq}$ (last row)
```

Table 6
Distributional parameters of simulated data sets ($K = 3$)

Balanced								Unbalanced							
Clusters' separation		Poor		Moderate		Good		Clusters' separation		Poor		Moderate		Good	
Cluster	Variable	μ_X	σ_X^2	μ_X	σ_X^2	μ_X	σ_X^2	Cluster	Variable	μ_X	σ_X^2	μ_X	σ_X^2	μ_X	σ_X^2
1 (30%)	X1	10.5	3.5	11.9	1.1	10.5	1.0	1 (60%)	X1	11.0	2.2	12.3	1.1	14.3	0.7
	X2	2.3	0.5	2.5	0.3	2.5	1.3		X2	5.3	0.8	6.4	0.6	7.0	0.2
	X3	7.8	2.0	8.0	0.9	4.3	1.8		X3	7.8	1.8	8.8	1.1	9.2	0.3
2 (30%)	X1	10.0	3.0	9.8	1.2	15.0	2.2	2 (30%)	X1	10.0	2.0	11.0	1.0	12.7	0.5
	X2	2.5	0.3	1.5	0.3	4.0	1.2		X2	4.5	0.5	5.0	0.5	5.0	0.4
	X3	7.0	1.0	6.8	0.7	7.0	1.5		X3	7.2	1.4	7.5	0.8	7.6	0.3
3 (40%)	X1	9.5	2.0	11.8	1.4	7.0	2.3	3 (10%)	X1	9.4	1.8	9.5	0.9	11.0	0.5
	X2	2.0	0.4	2.0	0.4	6.2	1.6		X2	4.0	0.4	3.7	0.4	3.5	0.3
	X3	7.5	1.2	8.9	0.7	2.5	1.7		X3	7.0	1.5	6.6	0.7	6.0	0.2
Average overlap		0.633		0.140		0.019		Average overlap		0.632		0.143		0.021	

If we specify $(1 - \alpha) = 0.99$ and $B = 0.01$ we obtain $T \geqslant \frac{1}{4}\left(\frac{2.576}{0.01}\right)^2$ which yields $T \geqslant 16{,}589.44$. Therefore, by generating 17,000 confusion matrices, we ensure the p-value (probability of an IPA, under H_0, is greater than the observed IPA value) is estimated with 99% confidence, [1]. It is worth noting that this 17,000 threshold is referred to the number of simulated values of indices and it applies to all confusion matrices regardless of their dimensions.

3.2. Experimental data to provide evaluation of IADJUST

In order to evaluate the performance of the IADJUST procedure, referring to indices in Table 3, several experimental scenarios are considered, based on finite mixtures of Gaussian distributed variables.

The following experimental factors are considered:

- Two, three or four clusters (with 2, 3 and 4 Gaussian attributes respectively, and 500, 800 and 1100 observations, respectively);
- Mixtures with balanced or unbalanced cluster weights or mixing proportions: balanced scenarios have equal or similar relative segment sizes; unbalanced scenarios have non-uniformly distributed cluster weights;
- Varying degrees of clusters overlapping with poor separation, moderate separation and good separation.

The different sample sizes are intended to prevent the small sample effect and feature space (over)dimensionality, not reflecting the full complexity of the underlying problem [17]. Thus, using the number of observations minus the number of parameters to be estimated as a proxy of degrees of freedom, the ratio between the "number of degrees of freedom" and the number of observations is close to 1. The distributional parameters of the simulated data sets with $K = 3$, for example, are presented in Table 6.

The R MixSim package [27], is used to obtain the simulated data and simultaneously control for the degree of overlap between clusters. The overlap measure $\omega_{kk'}$ quantifies the overlap between the k^{th} and k'^{th} components or clusters:

$$\omega_{kk'} = \omega_{k|k'} + \omega_{k'|k} \tag{27}$$

where $\omega_{k'|k}$ is the misclassification probability that one observation of a random vector originated from the k^{th} component is wrongly assigned to the k'^{th} component and $\omega_{\text{k}|k'}$ is defined similarly.

$$\omega_{k'|k} = P\left[\lambda_{k'}\phi\left(\underline{x};\underline{\mu}_{k'},\Sigma_{k'}\right) > \lambda_{k}\phi\left(\underline{x};\underline{\mu}_{k},\Sigma_{k}\right) \mid \underline{x} \sim N_p\left(\underline{\mu}_{k},\Sigma_{k}\right)\right] \tag{28}$$

where $\phi(\underline{x};\underline{\mu}_k,\Sigma_k)$ is a multivariate Gaussian density of the k^{th} component with mean vector $\underline{\mu}_k$, covariance matrix Σ_k and λ_k is the probability of occurrence of the k^{th} cluster.

According to the results reported [25], the degrees of separation of the clusters can be defined based on the $\omega_{kk'}$ measure: $\omega_{kk'} \approx 0.6$ for poorly-separated clusters; $\omega_{kk'} \approx 0.15$ for moderately-separated clusters; $\omega_{kk'} \approx 0.02$ for well-separated clusters.

Thirty data sets are generated in each of the 18 experimental scenarios amounting to 540 experimental data sets.

In order to illustrate the performance of IADJUST, we conduct the external validation of clustering results resorting to both the K-Means algorithm and the Expectation-Maximization (we use the Rmixmod package [23] for the estimation of a general Gaussian mixture model – $[P_K L_K C_K]$ in [6]). Clustering is based on the 540 generated data sets, yielding 1080 confusion (observed) matrices (EM and KM results).

4. Data analysis and results

4.1. The precision of IADJUST

Tables 7 and 8 display the results obtained by the proposed IADJUST method, along with the distributional or analytical-based results (DIST – for the Rand, Czekanowski, and Fowlkes and Mallows indices – and the approximation based results (APPROX) – for the Gower and Legendre, Jaccard and Sokal and Sneath (SS2) indices. These results refer to the confusion matrices obtained through the EM and KM clustering algorithms, when attempting to recover the structure of the data sets with three clusters; values are averaged over the 30 data sets and correspond to external validation of the clusters obtained. The indices of Goodman and Kruskal and *SoS* by Sokal and Sneath are excluded from these tables: *GK* needs no correction and IADJUST is the only available procedure for correction for *SoS*.

The high precision of IADJUST is evidenced in the perfect match between its results and the DIST results (with 3 decimals) for the Rand, Czekanowski, and Fowlkes and Mallows indices – in fact, the maximum error obtained is 0.000513, referring to the Fowlkes and Mallows index in the 4 Clusters/Poorly separated/Unbalanced scenario. APPROX method precision is reported in [3]: although the authors of this study do not give detailed information on the evaluation of the relationship between the Jaccard and Czekanowski indices, they report a maximum regression error of 0.0278. Note also that the present study's experimental scenarios are much wider in scope than the APPROX scenarios considered in [3], thus making the attainment of a smaller error more challenging.

The conclusions reported for the $K = 3$, in Tables 7 and 8 can be extended to the $K = 2$ and $K = 4$ scenarios, thus indicating that IADUST can determine very accurate E_o (*IPA*) values. The dispersion of IADJUST values depends mainly on clusters overlap: although it apparently decreases with the increase in overlap Fig. 1 the coefficient of variation (a measure of relative dispersion) is in fact higher within poorly separated clusters scenarios (around 0.5 for the results of 3 and 4 clusters and slightly surpassing 1 when 2 clusters are considered). For moderate and good separation, the average coefficient of variation is around 0.05 (all scenarios considered).

4.2. The detailed comparison of two clustering algorithms

Using IADJUST we were able to exclude agreement by chance in the external validation of EM and

Table 7
Average IPA values using the IADJUST, the DIST and the APPROX approaches (EM results for $K = 3$)

$K = 3$; EM	IPA	Clusters' separation								
		Poor			Moderate			Good		
		Iadjust	Distrib	Approx	Iadjust	Distrib	Approx	Iadjust	Distrib	Approx
Balanced	R	0.464	0.464	n.a.	0.521	0.521	n.a.	0.552	0.552	n.a.
	GL	0.632	n.a.	0.631	0.684	n.a.	0.687	0.711	n.a.	0.716
	J	0.275	n.a.	0.270	0.232	n.a.	0.224	0.204	n.a.	0.195
	C	0.431	0.431	n.a.	0.376	0.376	n.a.	0.339	0.339	n.a.
	SS2	0.160	n.a.	0.140	0.132	n.a.	0.106	0.114	n.a.	0.086
	FM	0.453	0.453	n.a.	0.381	0.381	n.a.	0.339	0.339	n.a.
Unbalanced	R	0.500	0.500	n.a.	0.505	0.505	n.a.	0.504	0.504	n.a.
	GL	0.666	n.a.	0.668	0.671	n.a.	0.673	0.670	n.a.	0.672
	J	0.313	n.a.	0.309	0.293	n.a.	0.289	0.296	n.a.	0.292
	C	0.476	0.476	n.a.	0.453	0.453	n.a.	0.457	0.457	n.a.
	SS2	0.186	n.a.	0.172	0.172	n.a.	0.155	0.174	n.a.	0.157
	FM	0.477	0.477	n.a.	0.454	0.454	n.a.	0.457	0.457	n.a.

Table 8
Average IPA values using the IADJUST, the DIST and the APPROX approaches (KM results for $K = 3$)

$K = 3$; KM	IPA	Clusters' separation								
		Poor			Moderate			Good		
		Iadjust	Distrib	Approx	Iadjust	Distrib	Approx	Iadjust	Distrib	Approx
Balanced	R	0.552	0.552	n.a.	0.551	0.551	n.a.	0.552	0.552	n.a.
	GL	0.711	n.a.	0.716	0.710	n.a.	0.715	0.711	n.a.	0.716
	J	0.204	n.a.	0.195	0.205	n.a.	0.196	0.204	n.a.	0.195
	C	0.339	0.339	n.a.	0.341	0.341	n.a.	0.339	0.339	n.a.
	SS2	0.114	n.a.	0.086	0.114	n.a.	0.087	0.114	n.a.	0.086
	FM	0.339	0.339	n.a.	0.341	0.341	n.a.	0.339	0.339	n.a.
Unbalanced	R	0.515	0.515	n.a.	0.514	0.514	n.a.	0.506	0.506	n.a.
	GL	0.680	n.a.	0.683	0.679	n.a.	0.681	0.672	n.a.	0.674
	J	0.239	n.a.	0.231	0.246	n.a.	0.238	0.285	n.a.	0.280
	C	0.386	0.386	n.a.	0.394	0.394	n.a.	0.443	0.443	n.a.
	SS2	0.136	n.a.	0.110	0.140	n.a.	0.115	0.166	n.a.	0.148
	FM	0.390	0.390	n.a.	0.398	0.398	n.a.	0.444	0.444	n.a.

KM clustering algorithms, thus providing a fair comparison between the two within the diverse simulated settings. Furthermore, the use of IADJUST provided an otherwise unavailable comparison between the performances of EM and KM, resorting to new adjusted indices (indices with no analytical correction) Fig. 1.

When considering clusters with good or moderate separation, the EM consistently evidences better performance. The Mann-Whitney tests find significant differences between the performance of EM and KM, with EM achieving the best results in external validation: the significance level is pre-set to $\alpha = 0.05$ and p-values range from 0.018 (for GK_a) to 0.024 ($SS2_a$). According to subsequent tests results within experimental scenarios, EM and KM perform equally within poor separation scenarios and also within balanced scenarios. Within the specific balanced-poor separation scenario KM performs better according to the R_a, C_a and SoS_a indices (the remaining indices find no significant differences).

The sensitivity of EM and KM to the experimental factors can be further explored using non-parametric tests:

- The performance of EM and KM both degrade with an increase in clusters overlap (significant associations according to Spearman correlation coefficients for all indices);

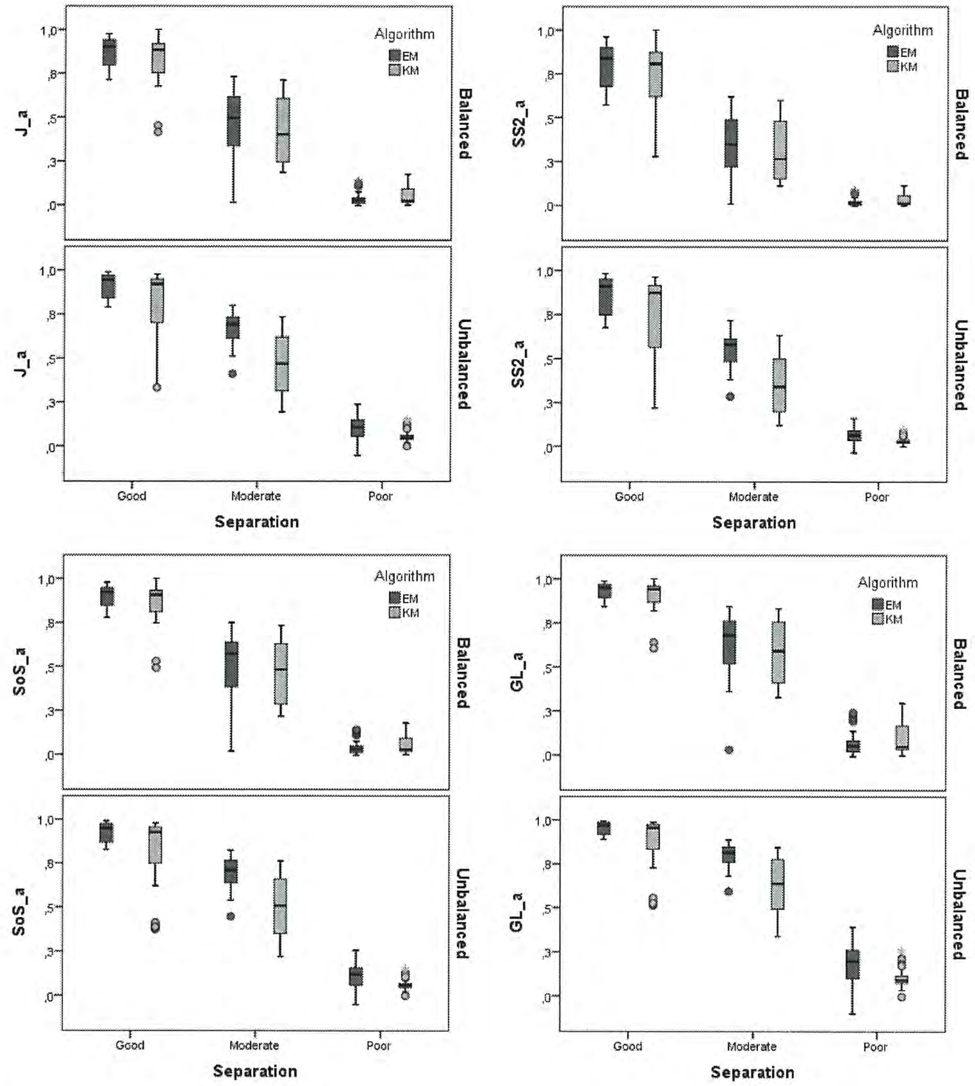

Fig. 1. The performance of EM vs. that of KM within experimental scenarios. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDA-150782)

- According to Mann-Whitney tests (for all indices), EM performs better in unbalanced than in balanced settings (significant differences); the same tests found no significant differences in the performance of KM between balanced and unbalanced settings.
- There is no significant association between EM's performance and the number of clusters according to most of the indices (except adjusted Jaccard and SS2); however, the performance of KM significantly decreases with the increase in the number of clusters (according to Spearman correlation coefficients for all indices).

Having concluded that EM attains the best results, we proceed the analysis with this algorithm's results in the experimental settings.

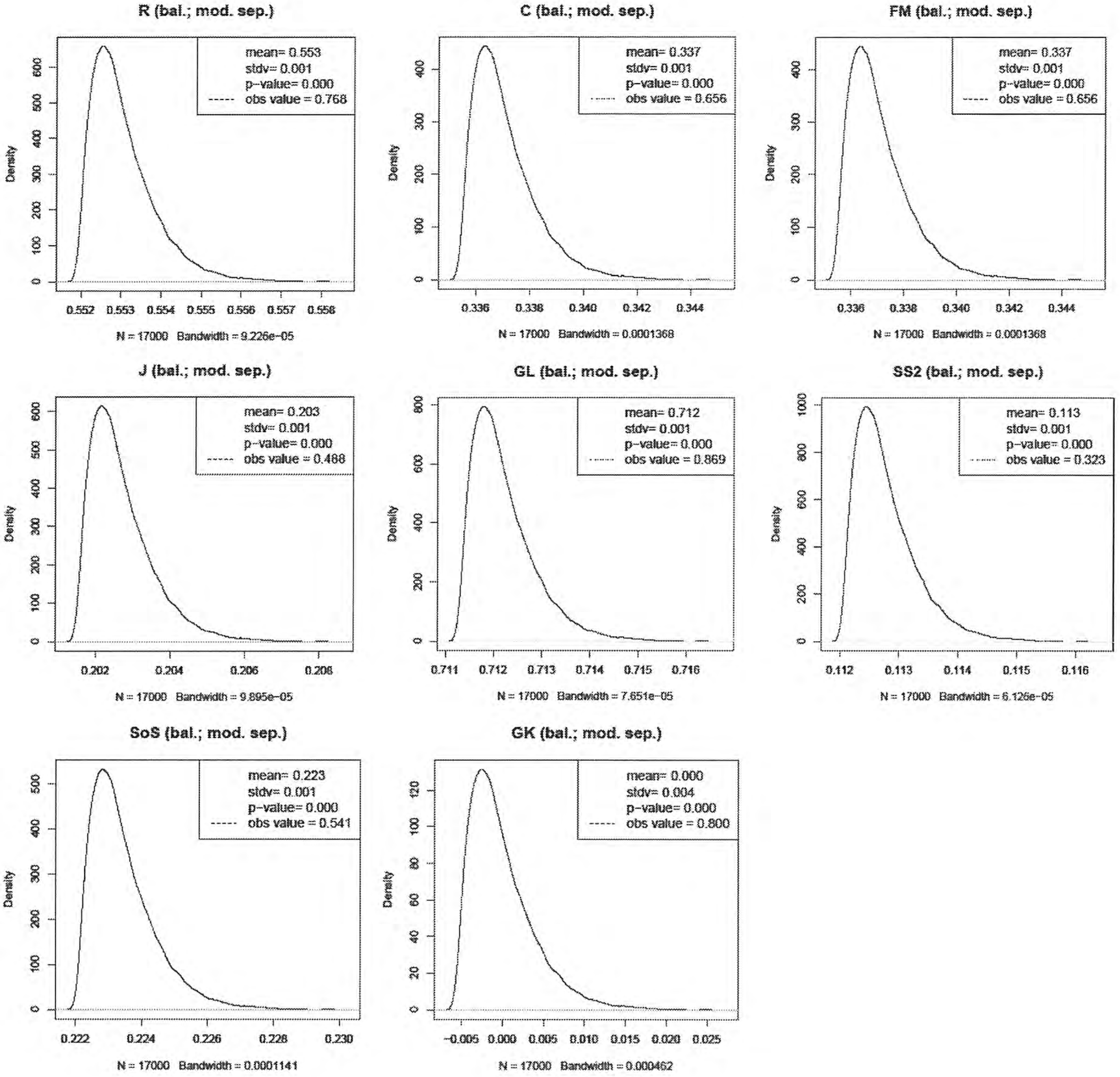

Fig. 2. Empirical distribution of paired indices of agreement under H_0 – scenario with balanced and moderately separated clusters, and $K = 3$ (EM results).

4.3. *New insights on paired indices of agreement*

4.3.1. *The indices' empirical distributions under the hypothesis of restricted agreement by chance*

Using the generated tables results (17,000 values for each of the 540 EM clustering results) we studied the empirical distributions of all indices under H_0. Figures 2 and 3 present two illustrative examples of the paired indices empirical distribution (under H_0) for $K = 3$. There is a clear distinction between the balanced and unbalanced scenarios: the former originate very skewed distributions (with positive skewness), while the later exhibit a much more symmetrical pattern; the unbalanced scenarios also induce

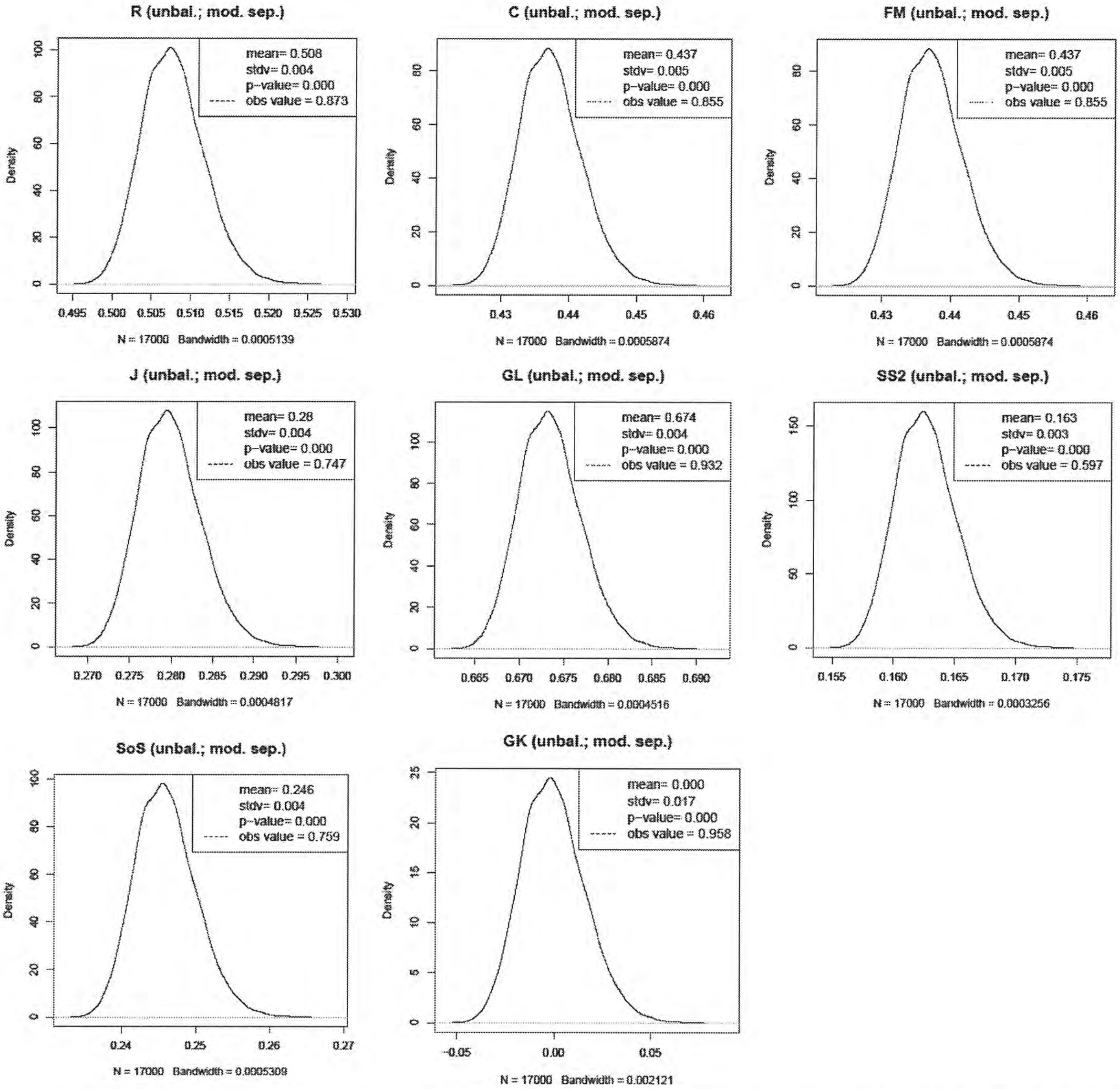

Fig. 3. Empirical distribution of paired indices of agreement under H_0 – scenario with unbalanced and moderately separated clusters, and $K = 3$ (EM results).

more dispersion in the indices values (significant differences, for $\alpha = 0.05$, are obtained with Mann-Whitney tests, for both skewness and range). This conclusion, for $K = 3$ clusters can be extended to the $K = 2$ and $K = 4$ scenarios, irrespective of the separation.

The Goodman and Kruskal index exhibits a "good behavior", particularly within the unbalanced scenario: average null values under H_0 and a symmetrical distribution. As already referred, this index needs no correction for agreement by chance and should therefore stand as a desirable pattern for the adjusted indices. On the other hand, it should be noted that, when compared to the remaining seven indices, this index has the largest dispersion under H_0 (highest range of values and highest coefficient of variation).

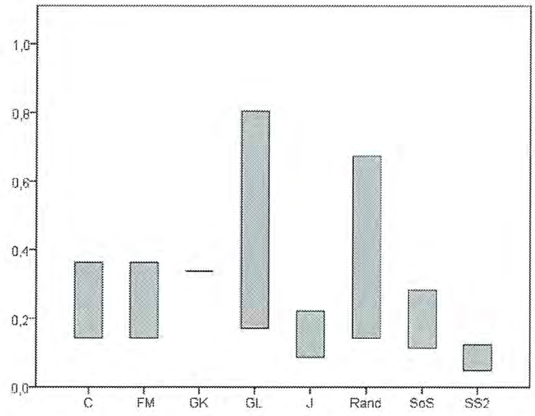

Fig. 4. The external validation of Statlog EM clustering – Difference between IOBSERV (bar top) and these values subtracted by their means under H_0 (bar bottom).

The Gower and Legendre index distribution is the most symmetrical (with the lowest skewness) but has a very high average value which is evidence of the need for correction.

4.3.2. The adjusted vs. the observed indices

When presenting the "threshold problem" for the Statlog data set, in the first section of our work, we drew attention to the need to fill a gap in the literature by providing adjusted versions for the various indices of paired agreement (we refer to eight indices in our work). We now revisit the Statlog problem and provide new insights regarding the adjustment of indices of paired agreement. In Fig 4, referring to the external validation of Statlog EM clustering, we illustrate the difference between the observed indices values (IOBSERV) and these values subtracted their average under H_0 (adjustment Eq. (1) numerator). After normalization, according to Eq. (1), IADJUST values tend to be more consistent while IOBSERV exhibit a marked variation. In particular, the Gower and Legendre, Sokal and Sneath (*SoS*) and Rand indices have the following values after correction: $GL_a = 0.17$, $SoS_a = 0.114$ and $R_a = 0.143$. In addition, the interpretation of adjusted indices is straightforward: they represent the recovery of the Statlog *a priori* known structure which goes beyond recovery by chance. Therefore, in the Statlog case, we conclude that recovery of the original structure is poor and very close to the null value that corresponds to agreement by chance. This same conclusion is derived from the eight indices under study, despite some small variations in the obtained IADJUST values.

The results obtained from synthetic data provide the means to compare the observed IPA values with these values subtracted by their means under H_0, within each experimental settings – see Fig. 5 for the 3 clusters scenario. Similar results are obtained for the solutions with 2 and 4 clusters.

The diverse IPA are affected differently by the $E_0\,(IPA)$ adjustment with the Gower and Legendre index being the most affected: the observed *GL* values are clearly misleading due to the incorporation of a large share of agreement by chance. As anticipated, the Goodman and Kruskal index does not need correction [8,37]. The Rand and the Czekanowski indices have similar adjustment patterns (in accordance with [2]), and very similar to the Fowlkes and Mallows corrected values – these 3 indices belong to the $\mathscr{L}$ family. The Sokal and Sneath indices have the lowest adjustment differentials.

Generally, the indices $E_0\,(IPA)$ adjustment – IADJUST numerator – increases with a decrease in separation (significant Spearman correlation coefficients, all above 0.887) and this is most relevant when considering poorly separated clusters. There is no clear distinction between unbalanced and balanced

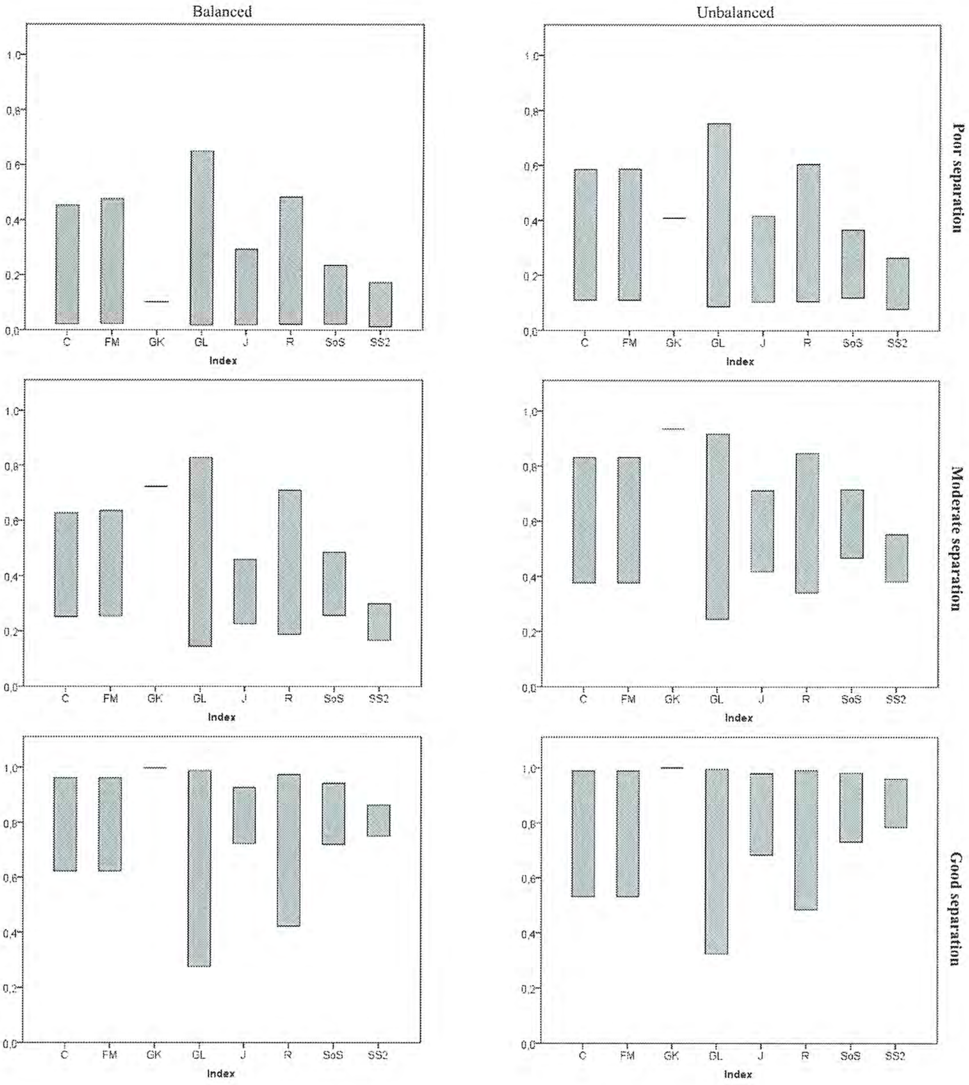

Fig. 5. The external validation of EM clustering of synthetic data with $K = 3$ – Difference between IOBSERV (bar top) and these values subtracted by their means under H_0 (bar bottom).

settings (the Mann-Whitney test is used): the J, C and FM indices show no significant differences between the two settings and the remaining indices are more adjusted in the balanced case. Moreover, the association between the adjustment gap and the number of clusters depends on the specific index considered: GK and Sokal and Sneath indices are unaffected by the number of clusters; the remaining indices are affected (correlation coefficients do significantly differ from zero).

The observed indices systematically exceed the 95% percentile of the corresponding distribution under H_0, which means that the agreement between recovered clustering structures and the *a priori* known classes significantly exceeds agreement by chance. There are, however, some exceptions, namely for the EM derived solutions and the poorly separated clusters scenario:

Table 9
Adjusted indices and experimental factors (measures of association values)

	K	Balance	Separation
	Spearman	Eta	Spearman
R_a	−0.077	0.137*	0.939*
GL_a	−0.079	0.143*	0.938*
J_a	−0.092*	0.147*	0.937*
C_a	−0.077	0.137*	0.939*
GK_a	−0.075	0.136*	0.941*
SoS_a	−0.075	0.131*	0.940*
SS2_a	−0.107*	0.154*	0.935*
FM_a	−0.076	0.137*	0.939*

*p-value < 0.05.

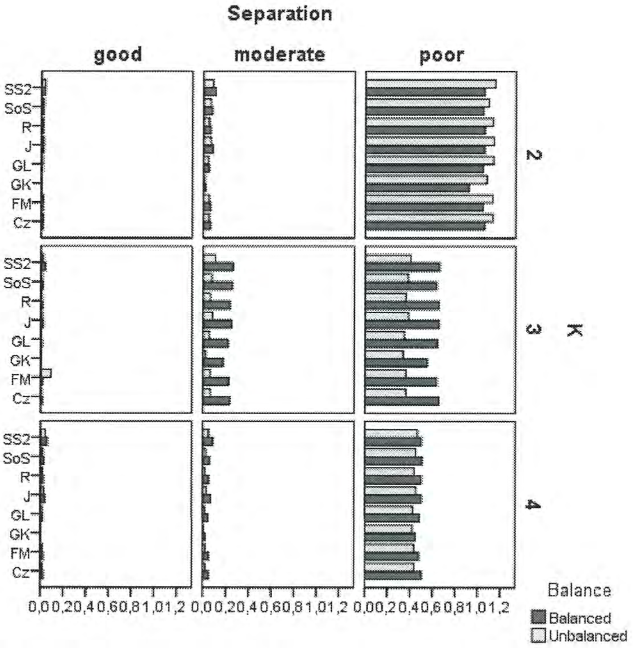

Fig. 6. Adjusted indices coefficient of variation within experimental scenarios.

- For $K = 2$ a third of all indices values are under the 95% percentile;
- For $K = 3$ a tenth of all indices values are under the 95% percentile;
- For $K = 4$ there is only one synthetic data set that originates a similar situation.

New insights are obtained after the indices' normalization, according to Eq. (1), namely referring to IADJUST values sensitivity to the experimental factors – Table 9. The indices values are clearly associated with the degree of separation: high positive Spearman correlation coefficients indicate that the greater the separation between clusters, the higher the indices (the better the EM performance). Adjusted indices are generally immune to the number of clusters, although tending to decrease slightly with the increase in this number. Adjusted indices are sensitive to balance: larger values are obtained in unbalanced settings.

When looking for a good index it is preferable to choose the one with the lower relative dispersion within experimental scenarios. In this context, the *GK* index appears to be the one having the lower coefficient of variation (ratio between standard deviation and mean) Fig. 6. In general, the indices relative dispersion clearly increases with the increase in overlap illustrating the corresponding clustering task difficulty.

5. Discussion and perspectives

In the present paper we focus on the correction of indices of paired agreement (IPA) between two partitions.

When comparing two partitions – e.g. when performing clustering validation and comparing estimated clusters and *real* clusters – agreement between them may be due to chance. This issue was first addressed when referring to a specific measure of paired agreement – the Rand index – leading to the proposal of a new adjusted Rand index [18]. Naturally, there are numerous IPA and this should be addressed when using any index. However, and despite some recent developments – e.g. [2,4,39] – the analytic correction cannot be provided for many indices. In addition, approximation approaches are limited in scope – e.g. [3] – and many indices do not yet have either an analytical nor approximate correction available.

The main contribution of this study is to provide a method that is able to correct any observed IPA for agreement by chance, thus overcoming limitations of previous studies. The proposed method – IADJUST – relies on the simulation of 17,000 confusion matrices under the hypothesis of restricted independence (H_0) and computes the corresponding indices averages under H_0.

Numerical experiments are conducted for eight IPA based on simulated data sets. We consider different scenarios for the data sets dimension (number of clusters in particular), the clusters overlap and balance. The R MixSim package is used for the generation of 540 data sets – [25].

The precision of IADJUST is illustrated by resorting to three indices with analytical solutions for correction (based on distributional assumptions): there is a negligible difference between IPA analytical averages and the averages provided by IADJUST thus highlighting the method's precision.

To illustrate the usefulness of the proposed method, we conduct external validation of the EM and KM algorithms within the diverse experimental scenarios. The diverse settings allow the pertinence of the indices correction to be analysed. We generally conclude that the observed indices overestimate the clustering performance, underlining the need to use adjusted indices (corrected for agreement by chance) when comparing two partitions, and notably when conducting external validation of clustering.

The detailed comparison between the EM and KM clustering algorithms, in the diverse simulated settings is an additional outcome of this study. The performance of the EM algorithm is generally better. EM and KM perform equally within poor separation scenarios and also within balanced scenarios. The performance of both EM and KM declines with an increase in clusters' overlap. While EM performs better in unbalanced than balanced settings (significant differences), no significant differences were found in the KM's performance between these settings.

After adjustment, the IPA belonging to the $\mathscr{L}$ family, including the well-known adjusted Rand index [18], yield identical values: they provide a similar perspective of the agreement between two partitions which suggests they capture the same common properties of clustering solutions. Therefore, it is advisable to ascertain the agreement between two partitions, including diverse IPA not restricted to this particular family. The IADJUST method, provides the adjustment of any index and thus offers an important contribution to widening the choice of possible indices for comparing two partitions.

This study revealed the empirical distributions, of the eight indices considered, under H_0. The results obtained highlight a significant difference in the skewness of these distributions between balanced and unbalanced scenarios: the former originate very skewed distributions (with positive skewness), while the later exhibit a much more symmetrical pattern; in addition, the unbalanced scenarios induce more dispersion in the indices values. This particular result questions the common practice of using the average index value, under H_0, for correction. Therefore, future research should test the use of the indices' percentiles (e.g. median), under H_0, as an alternative for correcting indices to exclude agreement by chance.

Future research should also consider further experimental scenarios – e.g. using additional numbers of clusters and dimensions and also addressing the validation of categorical clustering results.

References

[1] A. Agresti, D. Wackerly and J.M. Boyett, Exact conditional tests for cross-classifications: Approximation of attained significance levels, *Psychometrika* **44**(1) (1979), 75–83.

[2] A.N. Albatineh, Means and variances for a family of similarity indices used in cluster analysis, *Journal of Statistical Planning and Inference* **140** (2010), 2828–2838.

[3] A.N. Albatineh and M. Niewiadomska-Bugaj, Correcting jaccard and other similarity indices for chance agreement in cluster analysis, *Advances in Data Analysis and Classification* **5**(3) (2011), 179–200.

[4] A.N. Albatineh, M. Niewiadomska-Bugaj and D. Mihalko, On similarity indices and correction for chance agreement, *Journal of Classification* **23** (2006), 301–313.

[5] K. Bache and M. Lichman, *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science, (2013).

[6] C. Biernacki et al., Model-based cluster and discriminant analysis with the MIXMOD software, *Computational Statistics and Data Analysis* **51**(2) (2006), 587–600.

[7] R.L. Brennan and R.J. Light, Measuring agreement when two observers classify people into categories not defined in advance, *British Journal of Mathematical and Statistical Psychology* **27**(2) (1974), 154–163.

[8] N.J. Castellan, Jr., On the estimation of the tetrachoric correlation coefficient, *Psychometrika* **31**(1) (1966), 67–73.

[9] J. Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **51** (1960), 821–828.

[10] J. Czekanowski, Coefficient of racial likeness and durchschnittliche differenz, *Anthropologischer Anzeiger* **14** (1932), 227–249.

[11] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistics Society Series B (Methodological)* **39**(1) (1977), 1–38.

[12] E.B. Fowlkes and C.L. Mallows, A method for comparing two hierarchical clusterings, *Journal of the American Statistical Association* **78** (1983), 553–569.

[13] L.A. Goodman and W.H. Kruskal, Measures of association for cross classifications, *Journal of the American Statistical Associations* **49**(732–764) (1954).

[14] J.C. Gower and P. Legendre, Metric and euclidean Properties of dissimilarity coefficients, *Journal of Classification* **3** (1986), 5–48.

[15] J.H. Halton, A rigorous derivation of the exact contingency formula, *Proc Camb Phil Soc* **65** (1969), 527–530.

[16] J.A. Hartigan, *Clustering Algorithms*, N.Y.J.W.a. Sons, ed., 1975.

[17] T.K. Ho, Complexity of classification problems and comparative advantages of combined classifiers, in: *Multiple Classifier Systems*, Springer, (2000), 97–106.

[18] L. Hubert and P. Arabie, Comparing partitions, *Journal of Classification* **2**(1) (1985), 193–218.

[19] Jaccard, Nouvelles recerches sur la distribuition florale, *Bulletin De La Societé Vaudoise De Sciences Naturells* **44** (1908), 223–370.

[20] A.K. Jain, M.N. Murty and P.J. Flynn, Data clustering: A review, *ACM Computing Surveys (CSUR)* **31**(3) (1999), 264–323.

[21] M.G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, 2nd ed, Hafner, New York, 1967.

[22] W.J. Krzanowski and F.H.C. Marriott, *Multivariate Analysis*, Edward Arnold London, 1994.

[23] R. Lebret et al., *Rmixmod: The r Package of the Model-based Unsupervised, Supervised and Semi- Supervised Classification*, mixmod library, 2012.

[24] Z. Lu, Y. Peng and J. Xiao, From comparing clusterings to combining clusterings, in: *Twenty-Third AAAI Conference on Artificial Inteligence*, ed., (2008), 665–670.

[25] R. Maitra and V. Melnykov, Simulating data to study performance of finite mixture modeling and clustering algorithms, *Journal of Computational and Graphical Statistics* **19**(2) (2010), 354–376.

[26] P.D. McNicholas and T.B. Murphy, Model-based clustering of microarray expression data via latent gaussian mixture models, *Bioinformatics* **26**(21) (2010), 2705–2712.

[27] V. Melnykov, W.-C. Chen and R. Maitra, MixSim: An R package for simulating data to study performance of clustering algorithms, *Journal of Statistical Software* **51**(12) (2012), 1–25.

[28] V. Melnykov and I. Melnykov, Initializing the EM algorithm in gaussian mixture models with an unknown number of components, *Computational Statistics and Data Analysis* **56**(6) (2012), 1381–1395.

[29] G.W. Milligan and M.C. Cooper, A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavioral Research* **21**(4) (1986), 441–458.

[30] L.C. Morey and A. Agresti, The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement, *Educational and Psychological Measurement* **44**(1) (1984), 33–37.

[31] W. Patefield, Algorithm AS 159: An efficient method of generating random $R \times C$ tables with given row and column totals, *Applied Statistics* **30** (1981), 91–97.

[32] W.M. Rand, Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association* **66**(336) (1971), 846–850.

[33] R.R. Sokal and P.H. Sneath, *Principles of Numerical Taxonomy*, San Francisco CA:Freeman, 1963.

[34] D. Steinley, Properties of the hubert-arable adjusted rand index, *Psychological Methods* **9**(3) (2004), 386.

[35] D.L. Wallace, Comment on a method for comparing two hierarchical clusterings, *Journal of the American Statistical Association* **78** (1983), 569–576.

[36] M. Warrens, On the equivalence of cohen's kappa and the hubert-arabie adjusted rand index, *Journal of Classification* **25** (2008), 177–183.

[37] M.J. Warrens, On association coefficients for 2 × 2 tables and properties that do not depend on the marginal distributions, *Psychometrika* **73**(4) (2008), 777–789.

[38] M.J. Warrens, On similarity coefficients for 2 × 2 tables and correction for chance, *Psychometrika* **73**(3) (2008), 487–502.

[39] M.J. Warrens, On fixed points of the correction for chance function for 2 × 2 association coefficients, *International Journal of Research and Reviews in Applied Sciences* **15** (2013), 239–247.

[40] G.U. Yule, *An Introduction to the Theory of Statistics*, 6th ed, Charles Griffin and Company, 1922.

CHAPTER 5: CLUSTERING STABILITY AND GROUND TRUTH: NUMERICAL EXPERIMENTS

This manuscript has the following reference:

Amorim, M. J. & Cardoso, M. G. M. S. Clustering stability and ground truth: numerical experiments. 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) -, 2015a Lisboa. Science and Technology Publications, 259-264.

Clustering Stability and Ground Truth: Numerical Experiments

Maria José Amorim[1] and Margarida G. M. S. Cardoso[2]
[1]*Dep. of Mathematics, ISEL and Inst. Univ. de Lisboa (ISCTE-IUL), BRU-IUL, Av. Forças Armadas, Lisboa, Portugal*
[2]*Dep. of Quantitative Methods and BRU-UNIDE, ISCTE Busines School-IUL, Av. das Forças Armadas, Lisboa, Portugal*
mjamorim@adm.isel.pt, margarida.cardoso@iscte.pt

Keywords: Clustering, External Validation, Stability.

Abstract: Stability has been considered an important property for evaluating clustering solutions. Nevertheless, there are no conclusive studies on the relationship between this property and the capacity to recover clusters inherent to data ("ground truth"). This study focuses on this relationship resorting to synthetic data generated under diverse scenarios (controlling relevant factors). Stability is evaluated using a weighted cross-validation procedure. Indices of agreement (corrected for agreement by chance) are used both to assess stability and external validation. The results obtained reveal a new perspective so far not mentioned in the literature. Despite the clear relationship between stability and external validity when a broad range of scenarios is considered, within-scenarios conclusions deserve our special attention: faced with a specific clustering problem (as we do in practice), there is no significant relationship between stability and the ability to recover data clusters.

1 INTRODUCTION

Stability has been recognized as a desirable property of a clustering solution – e.g., (Jain and Dubes, 1988). A clustering solution is said to be stable if it remains fairly unchanged when the clustering process is subject to minor modifications such as alternative parameterizations of the algorithm used, introducing noise in the data or considering different samples. In order to evaluate stability, the agreement between the different clustering results originated by such minor modifications should be measured. Several indices of agreement (IA), such as the adjusted Rand (Hubert and Arabie, 1985), are commonly used for this end.

Some authors warn of a possible misuse of the property of clustering stability noting that the goodness of this property in the evaluation of clustering results is not theoretically well founded: "While it is a reasonable requirement that an algorithm should demonstrate stability in general, it is not obvious that, among several stable algorithms, the one which is most stable leads to the best performance" – ((Ben-David and Luxburg, 2008), p.1.) Bubeck et al., express a similar concern: "While model selection based on clustering stability is widely used in practice, its behavior is still not well-understood from a theoretical point of view"- ((Bubeck et al., 2012), p.436). Therefore, this study on clustering stability aims to contribute to clarify the role of this property in the evaluation of clustering results.

We focus on the relationship between clustering stability and its external validity i.e. agreement with "ground truth" – the true clusters' structures that are "a priori" known. Our aim is to obtain new insights based on diverse experimental scenarios.

We analyze diverse clustering results referred to 540 synthetic data sets generated under 18 different scenarios. Synthetic data sets provide straightforward clustering external evaluation and enable to control for diverse relevant factors such as the number of clusters, balance and overlapping – e.g., (Milligan and Cooper, 1985), (Vendramin et al., 2010), (Chiang and Mirkin, 2010).

2 THE PROPOSED METHOD

2.1 Why Stability?

Clustering stability, along with cohesion-separation, are commonly referred as a desirable properties of a clustering solution.

Cohesion-separation is intrinsically related with the concept of clustering and it can be related with the clusters' external validity - Milligan and Cooper

Amorim, M. and Cardoso, M..
Clustering Stability and Ground Truth: Numerical Experiments.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 259-264
ISBN: 978-989-758-158-8

(Milligan and Cooper, 1985) and Vendramin (Vendramin et al., 2010).

The value of stability is clearly related with the need to provide a useful clustering solution, since an inconsistent one would hardly serve practical purposes. On the other hand, the theoretical value of stability is yet to be understood.

Literature contributions on stability are discussed in Luxburg (Luxburg, 2009) and Ben-David and Luxburg (Ben-David and Luxburg, 2008), for example. These are specifically related with the capacity to recover the "right" number of clusters and to K-Means results. Another perspective of stability is offered in (Hennig, 2007) by measuring the consistency with which a particular cluster appears in replicated clustering - cluster-wise stability.

The lack of a systematical relationship between clusters validity and stability is occasionally pointed out by diverse studies - e.g., (Cardoso et al., 2010). Thus, a systematical study of the relationship between stability and clustering external validity is in order.

2.2 Cross-Validation

In order to evaluate clustering stability cross-validation can be used. Cross-validation referred to unsupervised analysis is described in (McIntyre and Blashfield, 1980).

In this work we resort to the weighted cross-validation procedure proposed in (Cardoso et al., 2010) to evaluate the stability of clustering solutions–see Table 1.

Table 1: Weighted cross-validation procedure.

Step	Action	Output
1	Perform training-test sample split	Weighted training and test samples
2	Cluster weighted training sample	Clusters in the weighted training sample
3	Cluster weighted test sample	Clusters in the weighted test sample
4	Obtain a contingency table between clusters obtained in 2. and 3.	Indices of d agreement values, indicators of stability

The “weighted training sample” considers unit weights for training observations (50% in the data sets considered) and almost zero weights to the remaining (test) observations. The “weighted test sample” reverses this weights’ allocation. The use of weighted samples overcomes the need for selecting a classifier when performing cross-validation. Furthermore, sample dimension is not a severe limitation for implementing clustering stability evaluation, since the Indices of agreement values are based on the entire (weighted) sample, and not in a holdout sample.

2.3 Adjusted Agreement between Partitions

In order to measure the agreement between two partitions we can resort to indices of agreement (IA).

In the literature, multiple IA can be found – e.g., (Vinh et al., 2010), (Warrens, 2008). They are generally quantified based on the cells values of the contingency table $[n_{kq}]$ between the two partitions P^K and P^Q being compared with K and Q clusters (respectively) - and on the corresponding row totals n_{k+} and column totals n_{+q}.

Among the IA, the Rand index ($Rand$) is, perhaps, the most well-known - (Rand, 1971).

$$Rand(P_I^K, P_{II}^K) = \frac{\binom{n}{2} + 2\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}} \quad (1)$$

It quantifies the proportion of all pairs of n observations that both partitions (P_I^K and P_{II}^K) agree to join in a group or to separate into different groups. Since agreement between partitions can occur by chance, (Hubert and Arabie, 1985) propose an adjusted version of $Rand$ using its expected value under the hypothesis of agreement by chance (H_o):

$$E_{H_0}\left[\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2}\right] = \frac{\sum_{k=1}^{K}\binom{n_{k+}}{2} \times \sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}} \quad (2)$$

Then this IA is adjusted according with the general formula:

$$IA_a(P^K, P^Q) = \frac{IA(P^K, P^Q) - E_{H0}[IA(P^K, P^Q)]}{Max[IA(P^K, P^Q)] - E_{H0}[IA(P^K, P^Q)]} \quad (3)$$

The adjusted index (IA_a) is thus null when agreement between partitions occurs by chance. Some IA are based on the concepts of entropy and information. Among these IA, Mutual Information (MI) is particularly well-known:

$$MI(P^K, P^Q) = \sum_{k=1}^{K}\sum_{q=1}^{Q}\frac{n_{kq}}{n}\log\left(\frac{n_{kq}}{\frac{n_{k+}n_{+q}}{n}}\right) \quad (4)$$

(Vinh et al., 2010) advocate a strategy similar to that of Hubert and Arabie, (Hubert and Arabie, 1985), to adjust MI for agreement by chance. These authors also advocate the use of a particular mutual information form resorting to joint entropy $H(P^K,P^Q)$ – ((Horibe, 1985), ((Kraskov et al., 2005)):

$$MIH(P^K,P^Q) = MI(P^K,P^Q)/H(P^K,P^Q) \qquad (5)$$

where

$$H(P^K,P^Q) = -\sum_{k=1}^{K}\sum_{q=1}^{Q}\frac{n_{kq}}{n}\log\left(\frac{n_{kq}}{n}\right) \qquad (6)$$

In this work we use the adjusted indices $Rand_a(P^K,P^Q)$ and $IMH_a(P^K,P^Q)$ to investigate agreement between two partitions. They offer different perspectives on agreement – paired agreement and simple agreement (Cardoso, 2007). These views are meant to provide useful insights when referring to external validation (comparison between the clustering solution and the "true" cluster structure) or to the evaluation of stability (comparison between two clustering solutions deriving from minor modifications in the clustering process).

3 NUMERICAL EXPERIMENTS

The pioneer study of Milligan and Cooper, (Milligan and Cooper, 1985), established the use of synthetic data to support the external validation of clustering structures. In this general setting, clustering solutions are to be compared with *a priori* known classes associated with the generated data sets. Since then, several works referring to external validation of clustering solutions have developed this line of work trying to overcome some drawbacks of this first study such as using the "right number of clusters" to quantify external validity is limited in scope, (Vendramin et al., 2010). Also, overlap between clusters should be properly quantified on the generation of experimental data sets (Steinley and Henson, 2005).

The present research considers three main design factors for the generation of synthetic data sets:

1. balanced (1- clusters are balanced having equal or very similar numbers of observations; 2- clusters are unbalanced)
2. number of clusters (K=2, 3,4)
3. clusters separation (1- poor; 2-moderate; 3- good)

The 18 resulting scenarios are named after the previous coding – for example, the scenario with balanced clusters (1), 3 clusters (3) and moderate separation (2) is termed "132".

The first design factor is operationalized as follows: balanced settings have classes with similar dimensions and for unbalanced settings classes have the following *a priori* probabilities or weights: a) 0.30 and 0.7 when K=2; b) 0.6, 0.3 and 0.1 when K=3; c) 0.5, 0.25, 0.15 and 0.10 when K=4.

The increasing number of clusters is associated with increasing number of variables (2, 3 and 4 latent groups with 2, 3 and 4 Gaussian distributed variables) and, in order to deal with this increasing complexity, we consider data sets with 500, 800 and 1100 observations, respectively.

The following measure of overlap between cluster is adopted, (Maitra and Melnykov, 2010):

$$\omega_{kk'} = \omega_{k|k'} + \omega_{k'|k} \qquad (7)$$

where $\omega_{k'|k}$ is the misclassification probability that the random variable X originated from the kth component is mistakenly assigned to the k'th component and $\omega_{k|k'}$ is defined similarly.

In order to generate the datasets within the scenarios, we capitalize on the recent contribution in (Maitra and Melnykov, 2010) and use the R MixSim package to generate structured data according to the finite Gaussian mixture model:

$$g(\underline{x}) = \sum_{k=1}^{K}\lambda_k\phi(\underline{x};\ \underline{\mu}_k, \Sigma_k) \qquad (8)$$

where $\phi(\underline{x};\ \underline{\mu}_k, \Sigma_k)$ is a multivariate Gaussian density of the kth component with mean vector $\underline{\mu}_k$ and covariance matrix Σ_k. Therefore

$$\omega_{k'|k} = P\left[\lambda_{k'}\phi\left(\underline{x};\ \underline{\mu}_{k'}, \Sigma_{k'}\right) > \lambda_k\phi\left(\underline{x}; \underline{\mu}_k, \Sigma_k\right) | \underline{x} \sim N_p\left(\underline{\mu}_k, \Sigma_k\right)\right] \qquad (9)$$

Based on this measure, we consider three degrees of overlap in the experimental scenarios: 1) $\omega_{kk'}$ is around 0.6 for poorly separated clusters; 2) $\omega_{kk'}$ is around 0.15 for moderately separated; 3) $\omega_{kk'}$ is around 0.02 for well separated classes (these thresholds are indicated in (Maitra and Melnykov, 2010)).

For each of the referred 18 scenarios, we generate 30 datasets and run our experiments by:

- clustering each data set;
- evaluating stability of the clustering solution (see 2.1 and 0);

- evaluating clustering external validity based on the *a priori* known classes (see 0);
- correlating results from stability and external validity to assess the role of the stability property.

The Rmixmod package is used for clustering purposes (Lebret et al., 2012). EM algorithm is found to be particularly suited for the clustering tasks at hand, since the data generated follow a finite Gaussian mixture model. We use the general Gaussian mixture model - [$P_K L_K C_K$] in (Biernacki et al., 2006).

The first results obtained are summarized in Table 2 and Table 3. They reveal the pertinence of the design factors, the overlap measure in particular: stability and external validity increase with the increase in separation, the IA being close to zero when separation is poor and near one when well separated clusters are considered. In general, the adjusted Rand index and normalized mutual information values illustrate the same underlying reality, although the $\mathrm{MIH_a}$ values provide a more conservative view of the degree of agreement between two partitions.

The general results referring to the relationship between stability and agreement with ground truth (inter experimental scenarios) are illustrated in Figure 1 and Figure 2. The corresponding Pearson correlation values are 0.958 and 0.933, respectively, indicating a high linear correlation between stability and external validity (both measured by MIH_a in Figure 1 and $Rand_a$ in Figure 2). These results corroborate the general theory on the relevance of the property of stability in the evaluation of clustering solutions.

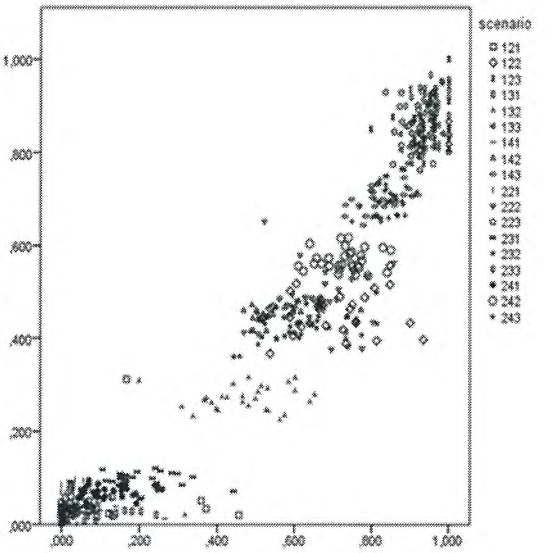

Figure 1: Inter-scenarios Pearson correlation between stability (yy') and agreement with ground truth (xx'): the $\mathrm{MIH_a}$ perspective.

Table 2: Adjusted Rand index values corresponding to external validity and to stability (values averaged over 30 datasets).

$\mathrm{Rand_a}$	Separation	External validity			Stability		
		K=2	K=3	K=4	K=2	K=3	K=4
Balanced	Poor	0.055	0.038	0.041	0.111	0.118	0.085
	Moderate	0.728	0.388	0.624	0.865	0.652	0.688
	Good	0.963	0.943	0.855	0.987	0.979	0.918
Unbalanced	Poor	0.097	0.211	0.133	0.053	0.280	0.166
	Moderate	0.765	0.690	0.820	0.864	0.822	0.898
	Good	0.962	0.980	0.887	0.981	0.991	0.949

Table 3: Normalized mutual information adjusted values corresponding to external validity and to stability (values averaged over 30 datasets).

$\mathrm{MIH_a}$	Separation	External validity			Stability		
		K=2	K=3	K=4	K=2	K=3	K=4
Balanced	Poor	0.046	0.024	0.031	0.073	0.054	0.073
	Moderate	0.458	0.263	0.449	0.700	0.465	0.578
	Good	0.865	0.832	0.707	0.949	0.931	0.833
Unbalanced	Poor	0.048	0.093	0.070	0.036	0.189	0.124
	Moderate	0.477	0.440	0.569	0.660	0.613	0.732
	Good	0.850	0.920	0.694	0.922	0.957	0.840

Table 4: Intra-scenarios Pearson correlations between stability and agreement with ground truth for synthetic data.

	Separation	MIH_a			Rand_a		
		K=2	K=3	K=4	K=2	K=3	K=4
Balanced	Poor	0.143	-0.018	-0.129	-0.079	-0.155	-0.303
	Moderate	0.122	0.264	-0.015	0.068	0.215	0.111
	Good	0.084	0.222	0.527	0.046	0.177	0.624
Unbalanced	Poor	0.329	0.126	0.172	0.367	-0.42	-0.079
	Moderate	-0.003	0.593	0.084	0.085	0.666	0.084
	Good	-0.151	0.272	0.245	-0.084	0.159	0.218

A completely different view is however provided intra-scenarios were very low correlations between stability and external validity are obtained – see Table 4. Within a specific scenario - the "real deal" for any clustering analysis practitioner - the correlation between external validity and stability is negligible. Both $Rand_a(P^K, P^Q)$ and $MIH_a(P^K, P^Q)$ lead to the same conclusion. Only two exceptions contradict this rule: scenarios "232" and "143".

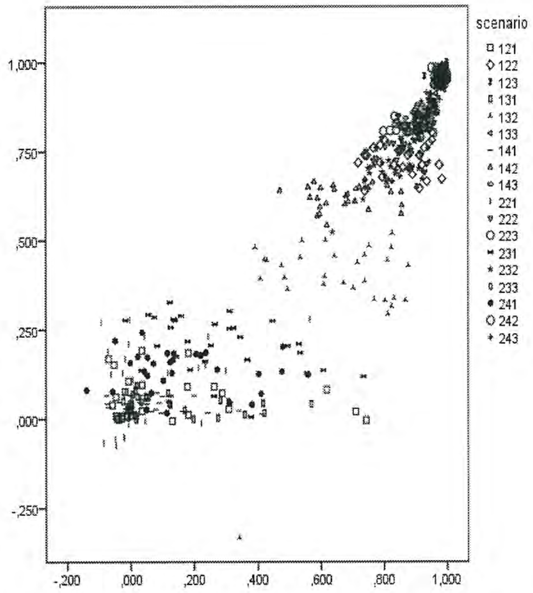

Figure 2: Inter-scenarios Pearson correlation between stability (yy') and agreement with ground truth (xx'): the Rand_a perspective.

4 CONTRIBUTIONS AND PERSPECTIVES

In this work we analyze the pertinence of using stability in the evaluation of a clustering solution. In particular, we question the following: does the consistency of a clustering solution (resisting minor modifications of the clustering process) provide indication towards a greater agreement with the "ground truth" (true structure) of the data?

In order to address this issue, we design an experiment in which 540 synthetic data sets are generated under 18 different scenarios. Design factors considered are the number of clusters, their balance and overlap. In addition, different sample sizes and space dimensions are considered.

Through the use of weighted cross-validation, we enable the analysis of stability, (Cardoso et al., 2010). We resort to adjusted indices of agreement (excluding agreement by chance) to measure agreement between two clustering solutions and also between a clustering solution and the "true" classes: we specifically use a simple index of agreement - the adjusted normalized Mutual Information, (Vinh et al., 2010) - and a paired one - the adjusted Rand índex (Hubert and Arabie, 1985).

A macro-view of the results does not contradict the current theory - there is a strong correlation between stability and external validity when the aggregate results are considered (all scenarios' results).

However, when it comes to perform clustering analysis within a specific experimental scenario, what can we say about the same correlation? The conclusions derived in this case support the previously referred concerns – there is an insignificant correlation between stability and external validity when it comes to a specific clustering problem.

Of course, it is still true that an unstable solution is, for this very reason, undesirable: then, which results should the practitioner consider? However, in a specific clustering setting, there is clearly no credible link between the stability of a partition and its approximation to ground truth.

This work contributes with a new perspective for a better understanding of the relationship between clustering stability and its external validity. To our knowledge, is the first time a study distinguishes between the macro view (all experimental scenarios considered) and the micro view (considering a specific clustering problem) and clearly differentiates the corresponding results.

In the future, stability results in discrete clustering should also be assessed and possible additional experimental factors also considered (e.g., the clusters' entropy).

In the future, clustering stability results in real data sets should also be assessed.

REFERENCES

Ben-David, S. & Luxburg, U. V., 2008. Relating clustering stability to properties of cluster boundaries. *In: Servedio, R. & Zhang, T., eds. 21st Annual Conference on Learning Theory (COLT), Berlin. Springer,* 379-390.

Biernacki, C., Celeux, G., Govaert, G. & Langrognet, F., 2006. Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis,* 51, 587-600.

Bubeck, S., Meila, M. & Von luxburg, U., 2012. How the initialization affects the stability of the k-means algorithm. *ESAIM: Probability and Statistics,* 16, 436-452.

Cardoso, M. G., Faceli, K. & De Carvalho, A. C., 2010. Evaluation of Clustering Results: The Trade-off Bias-Variability. *Classification as a Tool for Research. Springer*, 201-208.

Cardoso, M. G. M. S., 2007. Clustering and Cross-Validation. In: C. Ferreira, C. L., G. Saporta And M. Souto De Miranda, ed. IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal.

Celeux, G. & Diebolt, J., 1985. The SEM Algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly,* 2, 73-82.

Chiang, M. M.-T. & MIrkin, B., 2010. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification,* 27, 3-40.

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society. Series B (Methodological),* 39, 1-38.

Hartigan, J. A., 1975. *Clustering algorithms.*

Hennig, C., 2007. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis,* 52, 258-271.

Horibe, Y., 1985. Entropy and correlation. *Systems, Man and Cybernetics, IEEE Transactions on,* 5, 641-642.

Hubert, L. & Arabie, P., 1985. Comparing partitions. *Journal of Classification,* 2, 193-218.

Jain, A. K. & Dubes, R. C., 1988. *Algorithms for clustering data,* Englewood Cliffs, N.J.: Prentice Hall.

Kraskov, A., Stögbauer, H., Andrzejak, R. G. & Grassberger, P., 2005. Hierarchical clustering using mutual information. *EPL (Europhysics Letters),* 70, 278.

Lange, T., Roth, V., Braun, M. L. & Buchman, J. M., 2004. Stability based validation of clustering solutions. *Neural Computation,* 16, 1299-1323.

Lebret, R., S., L., Langrognet, F., Biernacki, C., Celeux, G. & Govaert, G., 2012. *Rmixmod: The r package of the model-based unsupervised, supervised and semi-supervised classification* [Online]. Rmixmod library. http://cran.rproject.org/web/packages/Rmixmod/index.html

Luxburg, U. V., 2009. Clustering Stability: An Overview. *Machine Learning,* 2, 235-274.

Maitra, R. & Melnykov, V., 2010. Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics,* 19, 354-376.

Mcintyre, R. M. & Blashfield, R. K., 1980. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research,* 2, 225-238.

Milligan, G. W. & Cooper, M. C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika,* 50, 159-179.

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association,* 66, 846-850.

Steinley, D. & Henson, R., 2005. OCLUS: an analytic method for generating clusters with known overlap. *Journal of Classification,* 22, 221-250.

Vendramin, L., Campello, R. J. & Hruschka, E. R., 2010. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining,* 3, 209-235.

Vinh, N. X., Epps, J. & Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research,* 11, 2837-2854.

Warrens, M. J., 2008. On similarity coefficients for 2× 2 tables and correction for chance. *Psychometrika,* 73, 487-502.

CHAPTER 6: CLUSTERING STABILITY AND GROUND TRUTH: NUMERICAL EXPERIMENTS

This manuscript has the following reference:

Amorim, M. J. & Cardoso, M. G. M. S. 2015b. Clustering stability and ground truth: numerical experiments. International Journal of Artificial Intelligence and Knowledge Discovery, to appear.

Clustering stability and ground truth: numerical experiments

Maria José Amorim
Mathematical Department
ISEL and ISCTE-IUL
Lisbon, Portugal
mjamorim@adm.isel.pt

Margarida G. M. S. Cardoso
Business Research Unit and Department of Quantitative
Methods for Management and Economics. ISCTE-IUL
Lisbon, Portugal
margarida.cardoso@iscte.pt

***Abstract*— Stability has been considered an important property for evaluating clustering solutions. Nevertheless, there are no conclusive studies on the relationship between this property and the capacity to recover clusters inherent to data ("ground truth"). This study focuses on this relationship, resorting to experiments on synthetic data generated under diverse scenarios (controlling relevant factors) and experiments on real data sets. Stability is evaluated using a weighted cross-validation procedure. Indices of agreement (corrected for agreement by chance) are used both to assess stability and external validation. The results obtained reveal a new perspective so far not mentioned in the literature. Despite the clear relationship between stability and external validity when a broad range of scenarios is considered, the within-scenarios conclusions deserve our special attention: faced with a specific clustering problem (as we do in practice), there is no significant relationship between clustering stability and the ability to recover data clusters**

Keywords- Clustering; external validation; stability.

I. INTRODUCTION

Stability has been recognized as a desirable property of a clustering solution – e.g. [1]. A clustering solution is said to be stable if it remains fairly unchanged when the clustering process is subject to minor modifications such as, alternative parameterizations of the algorithm used, introducing noise in the data or considering different samples. In order to evaluate stability, the agreement between the different clustering results originated by such minor modifications is measured. Several indices of agreement (IA), such as the adjusted Rand [2], are commonly used for this end.

Some authors warn of a possible misuse of the property of clustering stability noting that the goodness of this property in the evaluation of clustering results is not theoretically well founded: "While it is a reasonable requirement that an algorithm should demonstrate stability in general, it is not obvious that, among several stable algorithms, the one which is most stable leads to the best performance" – [3], p.1. Bubeck et al. express a similar concern: "While model selection based on clustering stability is widely used in practice, its behavior is still not well-understood from a theoretical point of view" - [4], p.436.

This study aims to contribute to clarify the role of stability in the evaluation of clustering results. We focus on the relationship between clustering stability and its external validity i.e. agreement with "ground truth" – the true clusters' structures that are "a priori" known.

In order to obtain new insights we consider diverse experimental scenarios and analyze diverse clustering results referred to 546 data sets. Synthetic data sets (540), generated under 18 different scenarios, provide straightforward clustering external evaluation and enable to control for diverse relevant factors such as the number of clusters, balance and overlapping – e.g. [5], [6], [7]. The use of 6 real data sets from the UCI Machine Learning Repository [8], complements the experimental analysis.

II. ON CLUSTERING STABILITY

A. Why stabilty?

Clustering stability, along with cohesion-separation, are commonly referred as desirable properties of a clustering solution. Cohesion-separation is intrinsically related with the concept of clustering and it can be related with the clusters' external validity - Milligan and Cooper [5] and Vendramin [6].

The value of stability is clearly related with the need to provide a useful clustering solution, since an inconsistent one would hardly serve practical purposes. On the other hand, the theoretical value of stability is yet to be understood.

Literature contributions on stability are discussed in Luxburg [9] and Ben-David and Luxburg [3], for example. These are specifically related with the capacity to recover the "right" number of clusters and to K-Means results. Another perspective of stability is offered in [10] by measuring the consistency with which a particular cluster appears in replicated clustering - cluster-wise stability.

The lack of a systematical relationship between clusters validity and stability is occasionally pointed out by diverse studies - e.g [11]. Thus, a systematical study of the relationship between stability and clustering external validity is in order.

B. Cross-Validation

In order to evaluate clustering stability cross-validation can be used. Cross-validation referred to unsupervised analysis, as described in [12], can be summarized into 5 main steps- Table 1.

TABLE 1. GENERAL CROSS-VALIDATION PROCEDURE

Step	Action	Output
1	Perform training-test Sample split	Training and test samples
2	Cluster training sample	Clusters in the training sample
3	Build a classifier using the training sample supervised by clusters' labels; use the classifier in the test sample.	Classes in the test sample
4	Cluster the test sample	Clusters in the test sample
5	Obtain a contingency table between clusters and classes in the test sample and calculate indices.	Indices of agreement values, indicators of stability

This clustering cross-validation procedure deserves, however, some remarks:

— Referring to step 3 [13] point out that "by selecting an inappropriate classifier, one can artificially increase the discrepancy between solutions (…) the identification of optimal classifiers by analytical means seems unattainable. Therefore, we have to resort to potentially suboptimal classifiers in practical applications", (p.1304-1305);

— In addition, the train-test split (step 1) requires sufficient sample size.

In this work, we resort to the weighted cross-validation procedure proposed in [11] to evaluate the stability of clustering solutions. The "weighted training sample" considers unit weights for training observations (50% in the data sets considered) and almost zero weights to the remaining (test) observations. The "weighted test sample" reverses this weights' allocation. The use of weighted samples overcomes the need for selecting a classifier when performing cross-validation. Furthermore, sample dimension is not a severe limitation for implementing clustering stability evaluation, since the Indices of agreement values are based on the entire (weighted) sample, and not in a holdout sample.

C. Adjusted agreement between partitions

In order to measure the agreement between two partitions we can resort to indices of agreement (*IA*). In the literature, multiple *IA* can be found – e.g. [14], [15]. They are generally quantified based on the cells values of the contingency table between the two partitions being compared - P^K and P^Q with K and Q clusters (respectively).

Among the *IA*, the Rand index (*Rand*) is, perhaps, the most well-known - [16].

$$Rand(P^K,P^Q) =$$

$$\frac{\binom{n}{2} + 2\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}} \qquad (1)$$

Where n_{kq} are the cells values of the contingency table, and n_{k+} and n_{+q} are the corresponding row totals and column totals, respectively.

It quantifies the proportion of pairs of observations that both partitions agree to join in a group or to separate into different groups. Since agreement between partitions can occur by chance, [2] propose an adjusted version of *Rand* using its expected value under the hypothesis of agreement by chance (H_o):

$$E_{H_o}\left[\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2}\right] = \frac{\sum_{k=1}^{K}\binom{n_{k+}}{2} \times \sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}}. \qquad (2)$$

Then this *IA* is adjusted according with the general formula:

$$IA_a(P^K,P^Q) =$$

$$\frac{IA(P^K,P^Q) - E_{H0}[IA(P^K,P^Q)]}{Max[IA(P^K,P^Q)] - E_{H0}[IA(P^K,P^Q)]}. \qquad (3)$$

The adjusted index (IA_a) is thus null when agreement between partitions occurs by chance. Some *IA* are based on the concepts of entropy and information. Among these *IA*, Mutual Information (*MI*) is particularly well-known:

$$MI(P^K,P^Q) = \sum_{k=1}^{K}\sum_{q=1}^{Q}\frac{n_{kq}}{n}\log\left(\frac{\frac{n_{kq}}{n}}{\frac{n_{k+}n_{+q}}{n}}\right). \qquad (4)$$

Vinh et al., [14], advocate a strategy similar to that of [2] to adjust *MI* for agreement by chance. These authors also advocate the use of a particular mutual information form resorting to joint entropy $H(P^K,P^Q)$ – ([17], [18]):

$$MIH(P^K,P^Q) = \frac{MI(P^K,P^Q)}{H(P^K,P^Q)}, \qquad (5)$$

where

$$H(P^K,P^Q) = -\sum_{k=1}^{K}\sum_{q=1}^{Q}\frac{n_{kq}}{n}\log\left(\frac{n_{kq}}{n}\right). \qquad (6)$$

In order to investigate agreement between two partitions we resort to the adjusted indices $Rand_a(P^K,P^Q)$ and $MIH_a(P^K,P^Q)$. They offer different perspectives on agreement – paired agreement and simple agreement [19]. These views are meant to provide useful insights when referring to external validation (comparison between the clustering solution and the "true" cluster structure) or to the evaluation of stability (comparison between two clustering solutions deriving from minor modifications in the clustering process).

III. Numerical Experiments

A. *Synthetic data*

The pioneer study of Milligan and Cooper, [5], established the use of synthetic data to support the external validation of clustering structures. In this general setting, clustering solutions are to be compared with *a priori* known classes associated with the generated data sets. Since then, several works referring to external validation of clustering solutions have developed this line of work trying to overcome some drawbacks of this first study such as using the "right number of clusters" to quantify external validity is limited in scope, [6]. In addition, overlap between clusters should be properly quantified on the generation of experimental data sets [20].

The present research considers three main design factors for the generation of synthetic data sets:

— balance (1- clusters are balanced having equal or very similar numbers of observations; 2- clusters are unbalanced)

— number of clusters (K=2, 3,4)

— clusters separation (1- poor; 2-moderate; 3- good).

The 18 resulting scenarios are named after the previous coding – for example, the scenario with balanced clusters (1), 3 clusters (3) and moderate separation (2) is termed "132".

The first design factor is operationalized as follows: balanced settings have classes with similar dimensions and for unbalanced settings classes have the following *a priori* probabilities or weights: a) 0.30 and 0.7 when K=2; b) 0.6, 0.3 and 0.1 when K=3; c) 0.5, 0.25, 0.15 and 0.10 when K=4.

The increasing number of clusters is associated with increasing number of variables (2, 3 and 4 latent groups with 2, 3 and 4 Gaussian distributed variables) and, in order to deal with this increasing complexity, we consider data sets with 500, 800 and 1100 observations, respectively.

The following measure of overlap between the classes *k* and *k'* is adopted, [21]:

$$\omega_{kk'} = \omega_{k|k'} + \omega_{k'|k} \,, \tag{7}$$

where $\omega_{k'|k}$ is the misclassification probability that the random variable *X* originated from the kth component is mistakenly assigned to the k'th component and $\omega_{k|k'}$ is defined similarly.

In order to generate the datasets within the scenarios, we capitalize on the recent contribution in [21] and use the R MixSim package to generate structured data according to the finite Gaussian mixture model:

$$\sum_{k=1}^{K} \lambda_k \phi(\underline{x};\, \underline{\mu}_k, \Sigma_k)\,, \tag{8}$$

where $\phi(\underline{x};\, \underline{\mu}_k, \Sigma_k)$ is a multivariate Gaussian density of the kth component with mean vector $\underline{\mu}_k$ and covariance matrix Σ_k. Therefore,

$$\begin{aligned}\omega_{k'|k} = P\Big[\lambda_{k'}\Phi\left(\underline{x};\, \underline{\mu}_{k'}, \Sigma_{k'}\right) > \\ \lambda_k\Phi\left(\underline{x};\underline{\mu}_k, \Sigma_k\right) | \underline{x} \sim N_p\left(\underline{\mu}_k, \Sigma_k\right)\Big].\end{aligned} \tag{9}$$

Based on this measure, we consider three degrees of overlap in the experimental scenarios: 1) $\omega_{kk'}$ is around 0.6 for poorly separated clusters; 2) $\omega_{kk'}$ is around 0.15 for moderately separated; 3) $\omega_{kk'}$ is around 0.02 for well separated classes. These thresholds are indicated in [21].

For each of the referred 18 scenarios, we generate 30 datasets and run our experiments by:

— clustering each data set;

— evaluating stability of the clustering solution (see II.A and II.C);

— evaluating clustering external validity based on the *a priori* known classes (see II.C);

— correlating results from stability and external validity to assess the role of the stability property.

The Rmixmod package is used for clustering purposes [22]. EM algorithm is found to be particularly suited for the clustering tasks at hand, since the data generated follow a finite Gaussian mixture model. We use the general Gaussian mixture model - [P$_K$L$_K$B$_K$] in [23].

The first results obtained are summarized in Table 2 and Table 3. They reveal the pertinence of the design factors: stability and external validity increase with the increase in separation, the *IA* being close to zero when separation is poor and near one when well separated clusters are considered. In general, the adjusted Rand index and mutual information values illustrate the same underlying reality, although the MIH_a values provide a more conservative view of the degree of agreement between two partitions.

The general results referring to the relationship between stability and agreement with ground truth (inter experimental scenarios), are illustrated in Figure1 and Figure 2. The corresponding Pearson correlation values are 0.958 and 0.933, respectively, indicating a high linear correlation between stability and external validity (both measured by MIH_a in Figure1 and $Rand_a$ in Figure 2. These results corroborate the general theory on the relevance of the property of stability in the evaluation of clustering solutions.

A completely different view is however provided intra-scenarios, yielding very low correlations between stability and external validity – see Table 4. Within a specific scenario - the "real deal" for any clustering analysis practitioner - the correlation between external validity and stability is negligible. Both the adjusted Rand and the adjusted Mutual Information lead to the same conclusion. Only two exceptions contradict this rule: scenarios "232" and "143".

B. *Real data*

The agreement between ground truth and stability is also subject to inspection in six data sets of the UCI Machine Learning Repository [8] – see Table 5 for a brief summary of these data sets. In addition to the design factors previously

TABLE 2 - ADJUSTED RAND INDEX VALUES CORRESPONDING TO EXTERNAL VALIDITY AND TO STABILITY (VALUES AVERAGED OVER 30 DATASETS).

$Rand_a$		External validity			Stability		
		K=2	K=3	K=4	K=2	K=3	K=4
Balanced	Poor	0.055	0.038	0.041	0.111	0.118	0.085
	Modar.	0.728	0.388	0.624	0.865	0.652	0.688
	Good	0.963	0.943	0.855	0.987	0.979	0.918
Unbalanced	Poor	0.097	0.211	0.133	0.053	0.280	0.166
	Modar.	0.765	0.690	0.820	0.864	0.822	0.898
	Good	0.962	0.980	0.887	0.981	0.991	0.949

TABLE 3 - MUTUAL INFORMATION ADJUSTED VALUES CORRESPONDING TO EXTERNAL VALIDITY AND TO STABILITY (VALUES AVERAGED OVER 30 DATASETS).

MIH_a		External validity			Stability		
		K=2	K=3	K=2	K=3	K=2	K=3
Balanced	Poor	0.046	0.024	0.031	0.073	0.054	0.073
	Modar.	0.458	0.263	0.449	0.700	0.465	0.578
	Good	0.865	0.832	0.707	0.949	0.931	0.833
Unbalanced	Poor	0.048	0.093	0.070	0.036	0.189	0.124
	Modar.	0.477	0.440	0.569	0.660	0.613	0.732
	Good	0.850	0.920	0.694	0.922	0.957	0.840

TABLE 4 - INTRA-SCENARIOS PEARSON CORRELATIONS BETWEEN STABILITY AND AGREEMENT FOR SYNTHETIC DATA

		MIH_a			$Rand_a$		
		K=2	K=3	K=4	K=2	K=3	K=4
Balanced	Poor	0.143	-0.018	-0.129	-0.079	-0.155	-0.303
	Mod.	0.122	0.264	-0.015	0.068	0.215	0.111
	Good	0.084	0.222	0.527	0.046	0.177	0.624
Unbalanced	Poor	0.329	0.126	0.172	0.367	-0.42	-0.079
	Mod.	-0.003	0.593	0.084	0.085	0.666	0.084
	Good	-0.151	0.272	0.245	-0.084	0.159	0.218

considered, we also quantify normalized entropy (ranging from 0 to 1 that indicates classes' uniform distribution).

Since the real data sets are diverse, we attempt to recover their clustering structures resorting to different clustering algorithms - namely the Hartigan K-Means (KM) algorithm [24], the Expectation Maximization (EM) [25] and the Stochastic EM (SEM) [26]. We resort to the EM and the SEM algorithms implemented in the Rmixmod package using the general Gaussian mixture model - $[P_K L_K B_K]$ in [23].

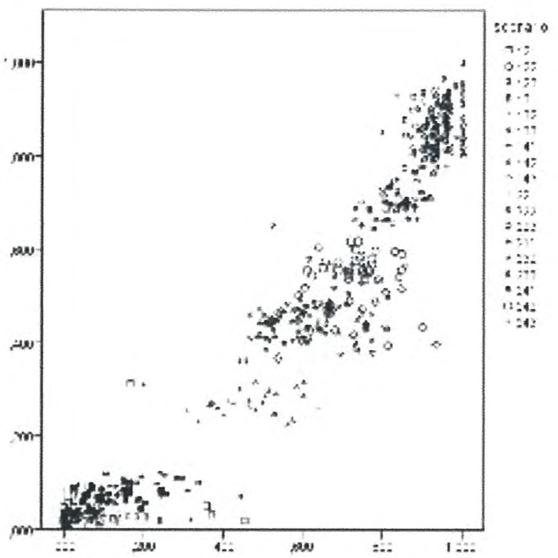

FIGURE1. INTER-SCENARIOS PEARSON CORRELATION BETWEEN STABILITY (YY') AND AGREEMENT WITH GROUND TRUTH (XX'): THE $MIH_A(P^K, P^Q)$ PERSPECTIVE

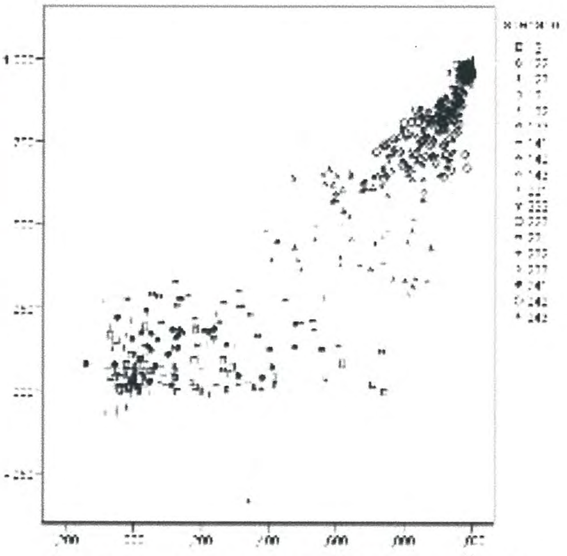

FIGURE 2. INTER-SCENARIOS PEARSON CORRELATION BETWEEN STABILITY (YY') AND AGREEMENT WITH GROUND TRUTH (XX'): THE $RAND_A(P^K, P^Q)$ PERSPECTIVE

TABLE 5 - REAL DATA SETS

Data set	n	Features	Classes	Normalized Entropy	Overlapping
Liver Disorders	345	6	C1 (145) C2 (200)	0.982	0.016
Wholesales	440	6	C1 (298) C2 (142)	0.907	0.111
Iris	150	4	Setosa (50) Versicolor (50) Virginica (50)	1.585	0.518
Wine recognition data	178	12	C1 (59) C2 (71) C3 (48)	1.567	0.002
Cars Silhouette	846	18	Bus (218) Saab (217) Opel (212) Van (199)	1.999067	0.044
User Modeling	258	5	Very-low (24) Low (83) Middle (88) High (63)	1.871	0.028

According to the results obtained (Table 6), the clustering solutions are generally stable, while agreement with ground truth varies appreciably. Thus, there is no relationship between stability and agreement with ground truth, the relationship under study appearing to be mainly dependent of the data set at hand.

IV. Contributions and Perspectives

In this work we analyze the pertinence of using stability in the evaluation of a clustering solution. In particular, we question the following: does the consistency of a clustering solution (resisting minor modifications of the clustering process) provide indication towards a greater agreement with the "ground truth" (true structure) of the data?

In order to address this issue, we design an experiment in which 540 synthetic data sets are generated under 18 different scenarios. Design factors considered are the number of clusters, their balance and overlap. In addition, different sample sizes and space dimensions are considered.

Through the use of weighted cross-validation, we enable the analysis of stability, [11]. We resort to adjusted indices of agreement (excluding agreement by chance) to measure agreement between two clustering solutions and also between a clustering solution and the "true" classes: we specifically use a simple index of agreement (IA) - the adjusted Mutual Information, [14] - and a paired IA - the adjusted Rand index [2].

A macro-view of the results does not contradict the current theory - there is a strong correlation between stability and external validity when the aggregate results are considered (all scenarios' results). However, when it comes to perform clustering analysis within a specific experimental scenario, what can we say about the same correlation? The conclusions derived in this study support the previously referred concerns referring to the relationship between stability and agreement with ground truth – there is an insignificant correlation between stability and external validity when it comes to a specific clustering problem.

TABLE 6.- STABILITY AND GROUND TRUTH FOR REAL DATA

Data set	Algo-rithm	Agreement with ground truth		Stability on Weighted train/test	
		Rand.	*MIH.*	*Rand.*	*MIH.*
Liver	KM	-0.005	-0.001	0.943	0.786
	EM	-0.009	0.002	0.960	0.844
	SEM	-0.010	0.002	0.987	0.933
Whole-sales	KM	0.564	0.311	-0.032	0.005
	EM	0.427	0.245	0.843	0.609
	SEM	0.427	0.251	0.851	0.621
Iris	KM	0.730	0.608	0.924	0.786
	EM	0.834	0.692	0.478	0.486
	SEM	0.834	0.699	0.478	0.486
Wine recogni-tion data	KM	0.352	0.264	0.760	0.615
	EM	0.915	0.805	0.802	0.691
	SEM	0.915	0.805	0.833	0.719
Cars	KM	0.126	0.099	0.651	0.552
	EM	0.143	0.102	0.601	0.521
	SEM	0.144	0.103	0.604	0.526
User Modeling	KM	0.189	-0.217	0.474	-0.126
	EM	0.372	0.118	0.574	0.245
	SEM	0.372	-0.131	0.531	-0.013

Of course, it is still true that an unstable solution is, for this very reason, undesirable (otherwise which results should the practitioner consider?). However, in a specific clustering setting, there is clearly no credible link between the stability of a partition and its approximation to ground truth.

This work contributes with a new perspective for a better understanding of the relationship between clustering stability and its external validity. To our knowledge, is the first time a study distinguishes between the macro view (all experimental scenarios considered) and the micro view (considering a specific clustering problem) and clearly differentiates the corresponding results.

In the future, stability results in discrete clustering should also be assessed and possible additional experimental factors (e.g. clusters' entropy) may also be considered.

References

1. Jain, A.K. and R.C. Dubes, Algorithms for clustering data. 1988: Englewood Cliffs, N.J.: Prentice Hall.
2. Hubert, L. and P. Arabie, "Comparing partitions", Journal of Classification, Vol 2, 1985, pp. 193-218.
3. Ben-David, S. and U.V. Luxburg, "Relating clustering stability to properties of cluster boundaries" in 21st Annual Conference on Learning Theory (COLT), Berlin: Springer, pp 379-390, July 2008.
4. Bubeck, S., M. Meila, and U. von Luxburg, "How the initialization affects the stability of the k-means algorithm" ESAIM: Probability and Statistics, Vol 16, 2012, pp. 436-452.

5. Milligan, G.W. and M.C. Cooper, "An examination of procedures for determining the number of clusters in a data set." Psychometrika, Vol 50, Issue 2, 1985, pp. 159-179.
6. Vendramin, L., R.J. Campello, and E.R. Hruschka, "Relative clustering validity criteria: A comparative overview" Statistical Analysis and Data Mining, Vol 3, Issue 4, 2010, pp. 209-235.
7. Chiang, M.M.-T. and B. Mirkin, "Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads" Journal of Classification, Vol 27, 2010, pp. 3-40.
8. Lichman, M. UCI Machine Learning Repository 2013; Available from: http://archive.ics.uci.edu/ml.
9. Luxburg, U.v., "Clustering Stability: An Overview" Machine Learning, Vol 2, issue 3, 2009, pp. 235-274.
10. Hennig, C., "Cluster-wise assessment of cluster stability" Computational Statistics & Data Analysis, Vol 52, 2007, pp. 258-271.
11. Cardoso, M.G., K. Faceli, and A.C. de Carvalho, Evaluation of Clustering Results: The Trade-off Bias-Variability, in Classification as a Tool for Research., Springer, pp. 201-208, 2010.
12. McIntyre, R.M. and R.K. Blashfield, "A nearest-centroid technique for evaluating the minimum-variance clustering procedure" Multivariate Behavioral Research, Vol 2, 1980, pp. 225-238.
13. Lange, T., et al., "Stability based validation of clustering solutions" Neural Computation, Vol 16, 2004, pp. 1299-1323.
14. Vinh, N.X., J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance" The Journal of Machine Learning Research, Vol 11, 2010, pp. 2837-2854.
15. Warrens, M.J., "On similarity coefficients for 2× 2 tables and correction for chance", Psychometrika, Vol 73, Issue 3, 2008, pp. 487-502.
16. Rand, W.M., "Objective criteria for the evaluation of clustering methods", Journal of the American Statistical Association, Vol 66, 1971, pp. 846-850.
17. Horibe, Y., "Entropy and correlation" Systems, Man and Cybernetics, IEEE Transactions, Vol 5, 1985, pp. 641-642.
18. Kraskov, A., et al., "Hierarchical clustering using mutual information", EPL (Europhysics Letters), Vol 70, Issue 2, 2005, pp. 278.
19. Cardoso, M.G.M.S. Clustering and Cross-Validation. in IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal, August 2007
20. Steinley, D. and R. Henson, "OCLUS: an analytic method for generating clusters with known overlap" Journal of Classification, Vol 22, Issue 2, 2005, pp. 221-250.
21. Maitra, R. and V. Melnykov, "Simulating data to study performance of finite mixture modeling and clustering algorithms" Journal of Computational and Graphical Statistics, Vol 19, 2010, pp. 354-376.
22. Lebret, R., et al." Rmixmod: The r package of the model-based unsupervised, supervised and semi- supervised classification" 2012; Available from: http://cran.r-project.org/web/packages/Rmixmod/ index.html
23. Biernacki, C., et al.," Model-Based Cluster and Discriminant Analysis with the MIXMOD Software" Computational Statistics and Data Analysis, Vol 51, 2006, pp. 587-600.
24. Hartigan, J.A., Clustering algorithms. 1975.
25. Dempster, A.P., N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm" Journal of the Royal Statistics Society. Series B (Methodological), Vol 39, 1977, pp. 1-38.
26. Celeux, G. and J. Diebolt, "The SEM Algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem" Computational Statistics Quarterly, Vol 2, 1985, pp. 73-8.

CHAPTER 7 - CONCLUSIONS AND FUTURE WORK

Conclusions and discussion

Data Analysis and Results

The analysis conducted first aims at providing adequate tools for the evaluation of clustering stability. The methodology proposed relies on the use of weighted cross-validation and resorts to several indices of agreement as indicators of stability.

In order to exclude agreement by chance when measuring the accordance between two partitions or crisp clustering solutions, a new method is proposed – IADJUST – that resorts to a simulation approach. This new approach overcomes limitations recognized in the literature and provides the correction of virtually any index of agreement based on cross-classification data.

The precision of IADJUST is illustrated by resorting to indices with known analytical solutions for correction – (Albatineh et al., 2006; Warrens, 2008b; Vinh et al., 2010a). We conclude that there is a negligible difference between these IA analytical averages and the averages provided by IADJUST– (Amorim and Cardoso, 2015c) and results of the example, in Table 8.

IADJUST is then used for correcting new indices and also to provide new insights on the indices distributions under the hypothesis of agreement by chance – (Amorim and Cardoso, 2014).

Through this contribution we make viable the correction of indices of agreement and clearly show the relevance of this adjustment, particularly for some indices – e.g. the NVI (simple) and the Gower & Legendre (paired) indices are the most affected by the adjustment. We thus advocate that the use of corrected indices of agreement should be a common practice, as opposed to the use of non-adjusted ones, not only when referring to clustering evaluation but also in the domain of consensus clustering.

What about stability

In this work we focus on the role of clustering stability in the evaluation of a clustering solution. For this end, we design an experiment in which 540 synthetic data sets are generated under 18 different scenarios. Design factors considered are the number of clusters, their balance and overlap.

Illustrating the usefulness of the IADJUST approach, we compare the stability of two clustering algorithms - (Amorim and Cardoso, 2012). According to the obtained results for the balanced data sets, the KM clustering results exhibit more stability then the EM results. However, the EM solutions are more stable when referring to the unbalanced data sets.

Finally, we address the relationship between the stability of a clustering solution and its external validity. A macro-view of the results obtained does not contradict the current theory since it shows there is a strong correlation between stability and external validity when the aggregate results are considered (all scenarios‟ results). However, within a specific experimental scenario (when a practical clustering task is considered), we find no relationship between stability and agreement with ground truth - (Amorim and Cardoso, 2015a). This is a new perspective on the relationship between clustering stability and its external validity. To our knowledge, is the first time a study distinguishes between the macro view (all experimental scenarios considered) and the micro view (considering a specific clustering problem). Nevertheless, and despite the results obtained, it is still true that an unstable solution is, for this very reason, undesirable: if not, which results should the practitioner consider?

This thesis work gave rise to a series of contributions that have materialized in several conferences presentations (Appendix D) as well as paper publications – Chaper 2, Chaper 3, Chaper 4, Chaper 5 and Chaper 6.

Future work

The investment in the IADJUST method should, in the near future, lead to make the software available on R, writing a new package, contributing to CRAN. This package should eventually extend the list of indices of agreement (simple and paired) considered in this work. Following the conclusions drawn about the indices distribution under the hypothesis of agreement by chance, the IADJUST procedure should be able to rely not only in the indices average values, but also in the median values (under H_0).

In future works, a now more informed selection of indices of agreement may offer different perspectives on the agreement between partitions and thus of clustering stability. A new typology of indices of agreement will be a natural outcome of the present study.

Future research should also address the stability of results in discrete clustering. And, when studying the relationship between stability and validity, additional experimental factors may also be considered (e.g., the “ground truth” clusters‟ entropy).

REFERENCES

Agresti, A., Wackerly, D. & Boyett, J. M. (1979), Exact conditional tests for cross-classifications: approximation of attained significance levels, *Psychometrika* 44, 75-83.

Albatineh, A. N. & Niewiadomska-Bugaj, M. (2011), Correcting Jaccard and other similarity indices for chance agreement in cluster analysis, *Adv. Data Anal. Classification* 5, 179-200.

Albatineh, A. N., Niewiadomska-BUgaj, M. & Mihalko, D. (2006), On Similarity Indices and Correction for Chance Agreement, *Journal of Classification* 23, 301-313.

Alizadeh, H., Mimaei-Bidgoli, B. & Parvin, H. (2014), To improve the quality of cluster ensembles by selecting a subset of base clusters, *Journal of Experimental and Theoretical Artificial Intelligence* 26, 127-150.

Amorim, M. J. & Cardoso, M. G. M. S. (2014), Paired Indices for clustering Evaluation. Correction for Agreement by Chance, *Proceedings of the 16th International Conference on Enterprise Information Systems*, Lisboa, Portugal, 164-170.

Amorim, M. J. & Cardoso, M. G. M. S. (2015a), Clustering stability and ground truth: numerical experiments, *Proceedings of the* 7th *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K),* Lisboa, Portugal, 259-264.

Amorim, M. J. & Cardoso, M. G. M. S. (2015b), Clustering stability and ground truth: numerical experiments, *International Journal of Artificial Intelligence and Knowledge Discovery*, to appear.

Amorim, M. J. & Cardoso, M. G. M. S. (2015c), Comparing clustering solutions: The use of adjusted paired indices, *Intelligent Data Analysis* 19, 1275–1296.

Amorim, M. J. P. C. & Cardoso, M. G. M. S.(2012), Clustering cross-validation and mutual information indices. *Proceedings of the 20th International Conference on Computational Statistics (COMPSTAT)*, Limassol, Cyprus, 39-52.

Bache, K. & Lichman, M. (2013) UCI Machine Learning Repository [Online]. Irvine, CA: University of California, School of Information and Computer Science, Available: http://archive.ics.uci.edu/ml.

Ben-David, S. & Luxburg, U. V. (2008), Relating clustering stability to properties of cluster boundaries. *Proceedings of the 21th Annual Conference on Learning Theory (COLT)*, Berlin, Germany, 379-390.

Ben-David, S., P´al, D. A. & Simon, H. U. (2007), Stability of k-Means Clustering *Proceedings of the 20th Annual Conference on Learning Theory (COLT)*, 20–34.

Ben-David, S., Simon, H. U. & P´al, D. A. A (2006), Sober Look at Clustering Stability. *Proceedings of the Annual Conference on Computational Learning Theory*, 5–19.

Ben-Hur, A., Elisseeff, A. & Guyon, I. (2002), A stability based method for discovering structure in clustered data, *Pacific Symposium on Biocomputing* 7, 6-17.

Biernacki, C., Celeux, G., Govaert, G. & Langrognet, F. (2006), Model-Based Cluster and Discriminant Analysis with the MIXMOD Software, *Computational Statistics and Data Analysis* 51, 587-600.

Breckenridge, J. (1989), Replicating Cluster Analysis: Method, Consistency and Validity, Multiv*ariate Behavioral Research* 24, 147-161

Cardoso, M. G. M. S. (2007), Clustering and Cross-Validation. *IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction*, 32 (CD-ROM).

Cardoso, M. G. M. S. & Carvalho, A. P. d. L. F. (2009), Quality Indices for (Practical) Clustering Evaluation, *Intelligent Data Analysis* 13, 725-740.

Cardoso, M.G., K. Faceli, & A.C. de Carvalho (2010), Evaluation of Clustering Results: The Trade-off Bias-Variability, *in Classification as a Tool for Research*, Springer, 201-208.

Cheng, R. & Milligan, G. W. (1996), Measuring the Influence of Individual. Data Points in a Cluster Analysis, *Journal of Classification* 13, 315-335.

Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 51, 821-828.

Czekanowski, J. (1932), "Coefficient of racial likeness" and "durchschnittliche Differenz", *Anthropologischer Anzeiger* 14, 227-249.

Dolnicar, S. & Leisch, F. (2010), Evaluation of structure and reproducibility of cluster solutions using the bootstrap, *Market Lett* 21, 83–101.

Dudoit, S. & Fridlyand, J. (2002), A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biology* 3, 1-21.

Everit, B., Landau, S. & Leese, M. (2001), *Cluster Analysis*. London. UK, Edward Arnold.

Fang, Y. & Wang, J. (2012), Selection of the number of clusters via the bootstrap method, *Computational Statistics and Data Analysis* 56, 468–477.

Fowlkes, E. B. & Mallows, C. L. (1983), A Method for Comparing Two Hierarchical Clusterings, *Journal of the American Statistical Association* 78, 553-569.

Fraley, C. & Raftery, A. E. (1998), How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, *The Computer Journal* 41, 578-588.

Fred, A. & Jain, A. K. (2003), Robust Data Clustering. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. USA.

Goodman, L. A. & Kruskal, W. H. (1954), Measures of Association for Cross Classifications, *Journal of the American Statistical Associations* 49, 732-764.

Gordon, A. D. (1999), *Classification*. Chapman & Hall/CRC.

Gower, J. C. & Legendre, P. (1986), Metric and Euclidean Properties of Dissimilarity Coefficients, *Journal of Classification* 3, 5-48.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001), On Clustering Validation Techniques, *Journal of Inteligent Information Systems* 17, 107-145.

Halton, J. H. (1969), A Rigorous Derivation of the Exact Contingency Formula, *Proc. Camb. Phil. Soc.* 65, 527-530.

Hartigan, J. A. (1975) Clustering algorithms.

Hennig, C. (2007), Cluster-wise assessment of cluster stability, *Computational Statistics & Data Analysis* 52, 258-271.

Hennig, C. & Liao, T. F. (2013), How to find an appropriate clustering for mixed type variables with application to socio-economic stratification, *Appl. Statist.* 62, 309–369.

Hubert, L. & arabie, P. (1985), Comparing partitions, *Journal of Classification* 2, 193-218.

Jaccard (1908), Nouvelles recerches sur la distribuition florale, *Bulletin de la Societé Vaudoise de Sciences Naturells* 44, 223-370.

Jain, A. K. & Dubes, R. C. (1988), *Algorithms for clustering data*. Englewood Cliffs, N.J.: Prentice Hall.

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999), Data Clustering: A Review. *ACM Computing Surveys* 31, 264-323.

Krey, S., Bratob, S., Liggesa, U., Götzeb, J. & Weihsa, C. (2015), Clustering of electrical transmission systems based on network topology and stability, *Journal of Statistical Computation and Simulation* 85, 47–61.

Krey, S., Ligges, U. & Leisch, F. (2014), Music and timbre segmentation by recursive constrained K-means clustering, *Computational Statistics* 29, 37-50.

Lange, T., Roth, V., Braun, M. L. & Buhmann, J. M. (2004), Stability-Based Validation of Clustering Solutions, *Neural Computation* 16, 1299–1323.

Luxburg, U. V. (2009), Clustering Stability: An Overview, *Machine Learning* 2, 235-274.

Maitra, R. & Melnykov, V. (2010), Simulating data to study performance of finite mixture modeling and clustering algorithms, *Journal of Computational and Graphical Statistics* 19 354-376.

Marateb, H. R., Mansourian, M., Adibi , P. & Farina, D. (2014), Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies, *Journal of Research in Medical Sciences* 19, 47-56.

Meilã, M. (2007), Comparing Clusterings - an information based distance, *Journal of Multivariate Analysis* 98, 873-895.

McIntyre, R.M. & Blashfield, R. K. (1980), A nearest-centroid technique for evaluating the minimum-variance clustering procedure, *Multivariate Behavioral Research* 2, 225-238.

Milligan, G. W. & Cooper, M. C. (1986), A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavioral Research* 21, 441-458.

Mirkin, B. (1996), *Mathematical Classification and Clustering*. Dordrecht /Boston/ London,Kluwer Academic Plublishers.

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* 52, 91-118.

Müller, H. & Hamm, U. (2014), Stability of market segmentation with cluster analysis – A methodological approach, *Food Quality and Preference* 34, 70-78.

Pascual, D., Pla , F. & Sánchez, J. S. (2010), Cluster validation using information stability measures, *Pattern Recognition Letters* 31, 454-461.

Patefield, W. M. (19819, Algorithm As159: An Efficient Method of Generating Random R * C Tables with Given Row and Column Totals. *Roayal Statistical Society, Series c, Applied Statistics* 30, 91-97.

Rand, W. M. (1971), Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association* 66, 846-850.

Ravera, O. (2001), A comparison between diversity, similarity and biotic indices applied to the macroinvertebrate community of a small stream: the Ravella river (Como Province, Northern Italy), *Aquatic Ecology* 35, 97–107.

Roth, V., Braun, M. L., Lange, T. & Buhmann, J. M. (2002), Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data. J.R. Dorronsoro (Ed.): ICANN 2002, LNCS 2415, 607–612.

Shamir, O. & Tishby, N. (2008a), Cluster stability for finite samples, *In: J. C. PLATT, D. K., Y. SInger & S. Roweis ed. Advances in neural information processing systems,* Cambridge:MIT Press., 1297–1304.

Shamir, O. & Tishby, N.(2008b), Model Selection and Stability in k-means clustering, *Proceedings of the 21th Annual Conference on learning theory*, Helsinki, Finland,. Cambridge: MIT Press, 67–378.

Shamir, O. & Tishby, N. (2010), Stability and model selection in k-means clustering. *Mach. Learn.* 80, 213-244.

Sokal, R. R. & Sneath, P. H. (1963), *Principles of Numerical Taxonomy*. San Francisco CA:Freeman.

Steinley, D. & Brusco, M. J. (2011), Choosing the Number of Clusters in K-Means Clustering, *Psychological Methods* 16, 285–297.

Strehl, A. & Gohosh, J. (2002), Cluster Ensembles- a Knowledge Reuse Framework for Combinig Partitions, *Journal of Machine Learning Research* 3, 583-617.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M. & Willett, P. (2012), Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets, *Journal of Chemical Information and Modeling* 52, 2884−2901.

Vendramin, L., Campello, R. J. & Hruschka, E. R. (2010), Relative clustering validity criteria: A comparative overview, *Statistical Analysis and Data Mining* 3, 209-235.

Vinh, N. X., Epps, J. & Bailey, J. (2010), Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, *The Journal of Machine Learning Research* 11, 2837-2854.

Wang, J. (2010), Consistent selection of the number of clusters via crossvalidation, *Biometrika* 97, 893–904.

Warrens, M. J. (2008), On association coefficients for 2*2 tables and properties that do not depend on the marginal distributions, *Psychometrika* 73, 777-789.

Warrens, M. J. (2008a), On Similarity coefficients for 2*2 tables and correction for chance, *Psycometrica* 73, 487-502.

Wu, Hang-Ming (2011), On biological validity indices for soft clustering algorithms for gene expression data, *Computational Statistics and Data Analysis* 55, 1969-1979.

APPENDIX A: THE EXACT ADJUSTMENT OF THE RAND INDEX

Under H_0, the probability, conditional on the row and column totals, of observing a specific cross-classification table (M) can be modeled by the Generalized Hypergeometric distribution (Halton, 1969),

$$f(M) = \frac{\prod_{i=1}^{K} n_{i\,.}! \prod_{j=1}^{Q} n_{.j}!}{n! \prod_{i=1}^{K} \prod_{j=1}^{Q} n_{ij}!} \qquad (5)$$

The total number of tables with identical row and column totals to the cross-classification in Table 6 a) can be easily listed considering that we have only one degree of freedom. For example, in order to keep the same marginal values the first element of the table must be between 40 and 10, change $n_{11} = 10\ to\ 40$ we get all possible tables. To calculate the average of the Rand index, we have to compute the Rand index value and the probability of each table (5). The results obtained are summarized in Table 11. The average is the sum of the values obtained by multiplying the index value by the probability.

Table 11 – List of all possible tables and respective Rand index and probability

Table		Rand	Probability	Table		Rand	Probability
40	10	0.778481	9.555E-14	24	26	0.494937	0.1644162
0	30			16	14		
39	11	0.741772	1.042E-11	23	27	0.498734	0.120357
1	29			17	13		
38	12	0.707595	4.912E-10	22	28	0.505063	0.0714023
2	28			18	12		
37	13	0.675949	1.34E-08	21	29	0.513924	0.0342109
3	27			19	11		
36	14	0.646835	2.391E-07	20	30	0.525316	0.0131712
4	26			20	10		
35	15	0.620253	2.984E-06	19	31	0.539241	0.0040464
5	25			21	9		
34	16	0.596203	2.719E-05	18	32	0.555696	0.0009829
6	24			22	8		
33	17	0.574684	0.0001865	17	33	0.574684	0.0001865
7	23			23	7		
32	18	0.555696	0.0009829	16	34	0.596203	2.719E-05
8	22			24	6		
31	19	0.539241	0.0040464	15	35	0.620253	2.984E-06
9	21			25	5		

Table 11- cont.

Table	Rand	Probability	Table	Rand	Probability
30 20 10 20	0.525316	0.0131712	14 36 26 4	0.646835	2.391E-07
29 21 11 19	0.513924	0.0342109	13 37 27 3	0.675949	1.34E-08
28 22 12 18	0.505063	0.0714023	12 38 28 2	0.707595	4.912E-10
27 23 13 17	0.498734	0.120357	11 39 29 1	0.741772	1.042E-11
26 24 14 16	0.494937	0.1644162	10 40 30 0	0.778481	9.555E-14
25 25 15 15	0.493671	0.1823924			

APPENDIX B: THE ADJUSTMENT OF THE RAND AND CZEKANWSKI INDICES

Considering the similarity matrix (counts of pairs of observations- Table 4:

$$a_{11} = \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2} = \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^{2} - \frac{n}{2} \tag{6}$$

$$a_{10} = \sum_{k=1}^{K}\binom{n_{k+}}{2} - a_{11} = \frac{1}{2}\sum_{k=1}^{K} n_{k+}^{2} - \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^{2} \tag{7}$$

$$a_{01} = \sum_{q=1}^{Q}\binom{n_{+q}}{2} - a_{11} = \frac{1}{2}\sum_{q=1}^{Q} n_{k+q\,.}^{2} - \frac{1}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K} n_{kq}^{2} \tag{8}$$

$$a_{01} = \binom{n}{2} - a_{11} - a_{10} - a_{01} \tag{9}$$

a_{11} - totals of pairs which are placed in the same clusters according to both partitions.

a_{00} - totals of pairs which are placed in different clusters according to both partitions.

$A = a_{11} + a_{00}$ - total numbers of agreements.

$$A = \binom{n}{2} + 2\sum\nolimits_{k=1}^{K}\sum\nolimits_{q=1}^{Q}\binom{n_{kq}}{2} - \left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right] \tag{10}$$

$a_{11} + a_{10} = \sum_{k=1}^{K}\binom{n_{k+}}{2}$ -totals of pairs which are placed in the same clusters of partition P^{k}.

$a_{11} + a_{01} = \sum_{q=1}^{Q}\binom{n_{+q}}{2}$ - totals of pairs which are placed in the same clusters of partition P^{Q}

$a_{10} + a_{00} = \binom{n}{2} - \sum_{q=1}^{Q}\binom{n_{+q}}{2}$ and $a_{01} + a_{00} = \binom{n}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2}$

Under the hypothesis of independence of the partitions being compared:

$$E(a_{11}) = \frac{(a_{11}+a_{10})(a_{11}+a_{01})}{a_{11}+a_{10}+a_{01}+a_{00}} = E\left(\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\right)$$

$$= \frac{\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}} \tag{11}$$

$$E(A) = \binom{n}{2} + 2E\left(\sum\nolimits_{k=1}^{K}\sum\nolimits_{q=1}^{Q}\binom{n_{kq}}{2}\right) - \left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]$$

$$= \binom{n}{2} + 2\frac{\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}} \tag{12}$$

$$-\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]$$

The adjustment of the Rand index

The Rand index (Table 5, p.9) is given by : $R = \frac{a_{11}+a_{10}}{a_{11}+a_{10}+a_{01}+a_{00}} = \frac{A}{\binom{n}{2}}$

$$E(R) = \frac{E(A)}{\binom{n}{2}}$$

$$= \frac{\binom{n}{2} + 2\frac{\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}} - \left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}}$$

$$= \frac{\binom{n}{2}}{\binom{n}{2}} + 2\frac{\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}^2} - \frac{\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}} \tag{13}$$

$$= \frac{\binom{n}{2}^2 + 2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} - \binom{n}{2}\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}^2}$$

$$= 1 + \frac{2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}^2} - \frac{\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}}$$

Equation (13) is Hubert and Arabie (1985 :198)“ equation 3.

The adjusted   index is $aR = \frac{R-E(R)}{1-E(R)}$, where:

$$
\begin{aligned}
&R - E(R)\\
&= \frac{\binom{n}{2} + 2\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - \left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}}\\
&- \frac{\binom{n}{2}^2 + 2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} - \binom{n}{2}\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}^2} =\\
&= \frac{\binom{n}{2}^2 + 2\binom{n}{2}\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - \binom{n}{2}\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right] - \binom{n}{2}^2}{\binom{n}{2}^2} \qquad (14)\\
&+ \frac{-2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} + \binom{n}{2}\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}^2}\\
&= \frac{2\binom{n}{2}\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - 2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}^2}
\end{aligned}
$$

using equation (14) the equation of aR can be rewritten:

$$
\mathrm{aR} = \frac{\dfrac{2\binom{n}{2}\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - 2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}^2}}{1 - 1 - \dfrac{2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}^2} + \dfrac{\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}{\binom{n}{2}}} \qquad (15)
$$

$$
= \frac{2\binom{n}{2}\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2} - 2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{-2\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} + \binom{n}{2}\left[\sum_{q=1}^{Q}\binom{n_{k+}}{2} + \sum_{k=1}^{K}\binom{n_{k+}}{2}\right]}
$$

Equation (15) is Hubert and Arabie (1985 :198)‟ equation 5.

The aR is equal to the $\mathbf{S_{Cohen}}$ index

The S_{Cohen} index, Cohen (1960), can be given by:

$$
S_{\text{Cohen}} = \frac{2(a_{11}a_{00} - a_{10}a_{01})}{(a_{11} + a_{10})(a_{10} + a_{00}) + (a_{11} + a_{01})(a_{01} + a_{00})} \qquad (16)
$$

where

$$\begin{aligned} a_{11}a_{00} &= \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\Bigg[\binom{n}{2} - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2} \\ &\quad - \sum_{k=1}^{K}\binom{n_{k+}}{2} + \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2} - \sum_{q=1}^{Q}\binom{n_{+q}}{2} \\ &\quad + \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\Bigg] = \binom{n}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2} \\ &\quad - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} \\ &\quad + \left(\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\right)^2 \end{aligned} \tag{17}$$

$$\begin{aligned} a_{10}a_{01} &= \left(\sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\right)\left(\sum_{q=1}^{Q}\binom{n_{+q}}{2} - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\right) \\ &= \sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} \\ &\quad - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} + \left(\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\right)^2 \end{aligned} \tag{18}$$

$$\begin{aligned}
&a_{11}a_{00} - a_{10}a_{01} \\
&\qquad = \binom{n}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2} \\
&\qquad - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} \\
&\qquad + \left(\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\right)^{2} \\
&\qquad - \sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} + \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} \\
&\qquad + \sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} - \left(\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2}\right)^{2} \\
&\qquad = \binom{n}{2}\sum_{q=1}^{Q}\sum_{k=1}^{K}\binom{n_{kq}}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}
\end{aligned} \tag{19}$$

$$\begin{aligned}
(a_{11} + a_{10})(a_{10} + a_{00}) &= \sum_{k=1}^{K}\binom{n_{k+}}{2}\left[\binom{n}{2} - \sum_{q=1}^{Q}\binom{n_{+q}}{2}\right] \\
&= \binom{n}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}
\end{aligned} \tag{20}$$

$$\begin{aligned}
(\mathrm{a}_{11} + a_{01})(a_{01} + a_{01}) &= \sum_{q=1}^{Q}\binom{n_{+q}}{2}\left[\binom{n}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2}\right] \\
&= \binom{n}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} - \sum_{q=1}^{Q}\binom{n_{+q}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2}
\end{aligned} \tag{21}$$

and

$$\begin{aligned}
&(a_{11} + a_{10})(a_{10} + a_{00}) + (\mathrm{a}_{11} + a_{01})(a_{01} + a_{01}) \\
&\qquad = \binom{n}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} - \binom{n}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} - 2\sum_{q=1}^{Q}\binom{n_{+q}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2}
\end{aligned} \tag{22}$$

Replace equations (17) to (22) in equation (16) we get:

$$\mathrm{S_{Cohen}} = \frac{2\left[\binom{n}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}\right]}{\binom{n}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2} - \binom{n}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2} - 2\sum_{q=1}^{Q}\binom{n_{+q}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2}}$$

$$= \frac{\sum_{k=1}^{K}\binom{n_{k+}}{2} - \frac{\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}}{\binom{n}{2}}}{\frac{1}{2}\left(\sum_{k=1}^{K}\binom{n_{k+}}{2} - \binom{n}{2}\sum_{q=1}^{Q}\binom{n_{+q}}{2}\right) - \frac{\sum_{q=1}^{Q}\binom{n_{+q}}{2}\sum_{k=1}^{K}\binom{n_{k+}}{2}}{\binom{n}{2}}} \qquad (23)$$

$$= \mathrm{aR}$$

The adjustment of the Czekanowski index

The Czekanowski index (Table 5) is given by: $C = \frac{2a_{11}}{2a_{11}+a_{10}+a_{01}} = \frac{2a_{11}}{(a_{11}+a_{10})+(a_{11}+a_{01})}$

Let $M = a_{11} + a_{10} + a_{01} + a_{00}$

$$E(C) = E\left(\frac{2a_{11}}{2a_{11} + a_{10} + a_{01}}\right) = E\left(\frac{2a_{11}}{(a_{11} + a_{10}) + (a_{11} + a_{01})}\right)$$

$$= \frac{2}{(a_{11} + a_{10}) + (a_{11} + a_{01})}E(a_{11}) \qquad (24)$$

$$= \frac{2}{(a_{11} + a_{10}) + (a_{11} + a_{01})} * \frac{(a_{11} + a_{10})(a_{11} + a_{01})}{M}$$

The aC is equal to the $\mathbf{S_{Cohen}}$ index

aC

$$= \frac{\frac{2a_{11}}{(a_{11} + a_{10}) + (a_{11} + a_{01})} - \frac{2}{(a_{11} + a_{10}) + (a_{11} + a_{01})} * \frac{(a_{11} + a_{10})(a_{11} + a_{01})}{M}}{1 - \frac{2}{(a_{11} + a_{10}) + (a_{11} + a_{01})} * \frac{(a_{11} + a_{10})(a_{11} + a_{01})}{M}}$$

$$= \frac{\frac{2a_{11}M - 2(a_{11} + a_{10})(a_{11} + a_{01})}{M * \left((a_{11} + a_{10}) + (a_{11} + a_{01})\right)}}{\frac{M * \left((a_{11} + a_{10}) + (a_{11} + a_{01})\right) - 2(a_{11} + a_{10})(a_{11} + a_{01})}{M * \left((a_{11} + a_{10}) + (a_{11} + a_{01})\right)}} \qquad (25)$$

$$= \frac{2(a_{11}a_{00} - a_{10}a_{01})}{2a_{11}a_{00} + a_{11}a_{10} + {a_{10}}^2 + a_{10}a_{00} + a_{11}a_{01} + {a_{01}}^2 + a_{01}a_{00}} = \mathrm{S_{Cohen}}$$

APPENDIX C –DATA ANALYSIS RESULTS

Since it was not possible to integrate all the results obtained in the published articles, we decided to make available some of them in this appendix.

Adjusted indices: comparative results

Table 12 - IPA expected values using the IADJUST, the DIST and the APPROX approaches (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=2; EM	IPA	Dataset - separation								
		Poor			Moderate			Good		
		iadjust	distrib	approx	iadjust	distrib	approx	iadjust	distrib	approx
Balanced	R	0.500	0.500		0.500	0.500		0.500	0.500	
	GL	0.667		0.668	0.667		0.668	0.667		0.668
	J	0.417		0.421	0.334		0.332	0.333		0.331
	Cz	0.586	0.586		0.500	0.500		0.500	0.500	
	SS2	0.265		0.273	0.200		0.191	0.200		0.190
	FM	0.600	0.600		0.500	0.500		0.500	0.500	
Unbalanced	R	0.528	0.528		0.516	0.516		0.512	0.512	
	GL	0.691		0.694	0.681		0.684	0.677		0.680
	J	0.447		0.452	0.417		0.421	0.404		0.408
	Cz	0.614	0.614		0.588	0.588		0.576	0.576	
	SS2	0.291		0.305	0.263		0.272	0.254		0.260
	FM	0.620	0.620		0.588	0.588		0.576	0.576	

Table 13 - IPA expected values using the IADJUST, the DIST and the APPROX approaches (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=2; KM	IPA	Dataset–separation								
		Poor			Moderate			Good		
		iadjust	distrib	approx	iadjust	distrib	approx	iadjust	distrib	approx
Balanced	R	0.500	0.500		0.500	0.500		0.500	0.500	
	GL	0.667		0.668	0.667		0.668	0.667		0.668
	J	0.333		0.332	0.333		0.332	0.333		0.331
	Cz	0.500	0.500		0.500	0.500		0.500	0.500	
	SS2	0.200		0.191	0.200		0.190	0.200		0.190
	FM	0.500	0.500		0.500	0.500		0.500	0.500	
Unbalanced	R	0.500	0.500		0.509	0.509		0.511	0.511	
	GL	0.667		0.669	0.674		0.677	0.676		0.679
	J	0.369		0.369	0.395		0.398	0.401		0.404
	Cz	0.539	0.539		0.566	0.566		0.572	0.572	
	SS2	0.226		0.224	0.246		0.250	0.251		0.256
	FM	0.540	0.540		0.566	0.566		0.572	0.572	

Table14 - IPA expected values using the IADJUST, the DIST and the APPROX approaches (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=4; EM	IPA	Dataset - separation								
		Poor			Moderate			Good		
		iadjust	distrib	approx	iadjust	distrib	approx	iadjust	distrib	approx
Balanced	R	0.515	0.515		0.623	0.623		0.625	0.625	
	GL	0.678		0.681	0.767		0.776	0.769		0.777
	J	0.193		0.183	0.144		0.135	0.143		0.134
	Cz	0.323	0.323		0.252	0.252		0.250	0.250	
	SS2	0.107		0.079	0.078		0.053	0.077		0.053
	FM	0.340	0.340		0.252	0.252		0.250	0.250	
Unbalanced	R	0.498	0.498		0.547	0.547		0.548	0.548	
	GL	0.664		0.666	0.708		0.712	0.708		0.713
	J	0.255		0.249	0.210		0.201	0.209		0.200
	Cz	0.406	0.406		0.346	0.346		0.345	0.345	
	SS2	0.147		0.124	0.117		0.089	0.116		0.089
	FM	0.415	0.416		0.346	0.346		0.345	0.345	

Table 15 - IPA expected values using the IADJUST, the DIST and the APPROX approaches (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=4; KM	IPA	Dataset–separation								
		Poor			Moderate			Good		
		iadjust	distrib	approx	iadjust	distrib	approx	iadjust	distrib	approx
Balanced	R	0.624	0.624		0.616	0.616		0.622	0.622	
	GL	0.769		0.777	0.762		0.770	0.767		0.775
	J	0.143		0.134	0.148		0.139	0.144		0.135
	Cz	0.251	0.251		0.258	0.258		0.252	0.252	
	SS2	0.077		0.053	0.080		0.055	0.078		0.053
	FM	0.251	0.251		0.259	0.259		0.252	0.252	
Unbalanced	R	0.577	0.577		0.565	0.565		0.559	0.559	
	GL	0.732		0.738	0.722		0.728	0.717		0.722
	J	0.170		0.161	0.187		0.177	0.196		0.186
	Cz	0.291	0.291		0.315	0.315		0.327	0.327	
	SS2	0.093		0.066	0.103		0.075	0.108		0.080
	FM	0.295	0.295		0.316	0.316		0.328	0.328	

Empirical results

The indices’ empirical distributions under the hypothesis of restricted agreement by chance

Table16 – The indices under H0: descriptive statistics

K=2; EM solutions		Clusters‟ separation											
		Poor				Moderate				Good			
		Min	Max	Ass	desvio	Min	Max	Ass	desvio	Min	Max	Ass	desvio
Balanced	Rand	0.499	0.510	8.587	0.001	0.499	0.516	11.972	0.001	0.499	0.516	11.684	0.001
	GL	0.666	0.676	8.489	0.001	0.666	0.681	11.778	0.001	0.666	0.681	11.505	0.001
	J	0.416	0.425	8.680	0.001	0.333	0.348	12.172	0.001	0.332	0.348	11.867	0.001
	Cz	0.585	0.595	8.587	0.001	0.499	0.517	11.972	0.001	0.499	0.516	11.684	0.001
	GK	-0.006	0.057	8.580	0.006	-0.004	0.065	11.954	0.006	-0.004	0.064	11.669	0.006
	SoS	0.201	0.214	8.736	0.001	0.249	0.267	12.270	0.001	0.249	0.266	11.957	0.001
	SS2	0.264	0.271	8.760	0.001	0.200	0.211	12.337	0.001	0.199	0.210	12.017	0.001
	FM	0.599	0.609	8.587	0.001	0.499	0.517	11.972	0.001	0.499	0.516	11.684	0.001
Unbalanced	Rand	0.516	0.555	2.568	0.005	0.501	0.554	0.704	0.007	0.500	0.547	0.928	0.006
	GL	0.680	0.713	2.498	0.004	0.667	0.713	0.630	0.006	0.666	0.707	0.851	0.005
	J	0.437	0.470	2.637	0.004	0.404	0.450	0.777	0.006	0.394	0.435	1.004	0.005
	Cz	0.605	0.636	2.568	0.004	0.575	0.620	0.704	0.006	0.565	0.606	0.928	0.005
	GK	-0.099	0.184	2.526	0.046	-0.067	0.160	0.676	0.028	-0.053	0.145	0.905	0.025
	SoS	0.192	0.245	2.677	0.007	0.225	0.284	0.821	0.007	0.231	0.282	1.049	0.006
	SS2	0.282	0.310	2.699	0.004	0.253	0.290	0.846	0.004	0.245	0.278	1.075	0.004
	FM	0.611	0.642	2.568	0.004	0.575	0.620	0.704	0.006	0.565	0.606	0.928	0.005

Table 17 – The indices under H0: descriptive statistics

		Clusters‟ separation											
		Poor				Moderate				Good			
K=3; EM solutions		Min	Max	Skew	Stdv	Min	Max	Skew	Stdv	Min	Max	Skew	Stdv
Balanced	Rand	0.460	0.473	1.761	0.002	0.519	0.530	2.417	0.001	0.550	0.558	3.192	0.001
	GL	0.628	0.640	1.733	0.001	0.682	0.692	2.386	0.001	0.710	0.717	3.163	0.001
	J	0.272	0.284	1.792	0.001	0.230	0.241	2.455	0.001	0.203	0.212	3.233	0.001
	Cz	0.427	0.441	1.761	0.002	0.373	0.387	2.417	0.002	0.338	0.349	3.192	0.001
	GK	-0.025	0.055	1.768	0.010	-0.012	0.042	2.405	0.006	-0.006	0.033	3.168	0.004
	SoS	0.208	0.223	1.805	0.002	0.225	0.238	2.462	0.001	0.223	0.233	3.234	0.001
	SS2	0.158	0.166	1.815	0.001	0.130	0.137	2.483	0.001	0.113	0.118	3.261	0.001
	FM	0.449	0.464	1.761	0.002	0.379	0.393	2.417	0.002	0.338	0.349	3.192	0.001
Unbalanced	Rand	0.489	0.519	0.638	0.004	0.491	0.526	0.245	0.004	0.490	0.527	0.223	0.005
	GL	0.657	0.683	0.609	0.003	0.659	0.689	0.219	0.004	0.657	0.690	0.196	0.004
	J	0.303	0.330	0.670	0.003	0.281	0.313	0.275	0.004	0.283	0.317	0.254	0.004
	Cz	0.464	0.495	0.638	0.004	0.439	0.477	0.245	0.005	0.441	0.482	0.223	0.005
	GK	-0.046	0.078	0.636	0.015	-0.053	0.086	0.241	0.017	-0.059	0.092	0.218	0.018
	SoS	0.235	0.266	0.684	0.004	0.234	0.270	0.286	0.004	0.233	0.272	0.266	0.005
	SS2	0.179	0.198	0.695	0.002	0.164	0.186	0.299	0.003	0.165	0.189	0.279	0.003
	FM	0.466	0.497	0.638	0.004	0.439	0.477	0.245	0.005	0.441	0.482	0.223	0.005

Table 18 – The indices under H0: descriptive statistics

		Clusters‟ separation											
		Poor				Moderate				Good			
K=4; EM solutions		Min	Max	Skew	Stdv	Min	Max	Skew	Stdv	Min	Max	Skew	Stdv
Balanced	Rand	0.514	0.519	2.694	0.001	0.622	0.626	1.372	0.000	0.624	0.628	1.342	0.000
	GL	0.678	0.682	2.678	0.000	0.767	0.770	1.364	0.000	0.768	0.771	1.334	0.000
	J	0.192	0.197	2.715	0.001	0.143	0.149	1.388	0.001	0.142	0.147	1.358	0.001
	Cz	0.321	0.328	2.694	0.001	0.250	0.259	1.372	0.001	0.248	0.256	1.342	0.001
	GK	-0.005	0.023	2.689	0.003	-0.007	0.022	1.349	0.003	-0.007	0.022	1.318	0.003
	SoS	0.210	0.216	2.718	0.001	0.187	0.194	1.383	0.001	0.186	0.193	1.353	0.001
	SS2	0.106	0.109	2.729	0.000	0.077	0.080	1.398	0.000	0.076	0.079	1.367	0.000
	FM	0.339	0.347	2.694	0.001	0.250	0.259	1.372	0.001	0.248	0.256	1.342	0.001
Unbalanced	Rand	0.485	0.516	0.121	0.004	0.537	0.563	0.160	0.003	0.538	0.563	0.136	0.003
	GL	0.652	0.680	0.106	0.004	0.699	0.720	0.144	0.003	0.699	0.721	0.120	0.003
	J	0.243	0.272	0.139	0.004	0.199	0.226	0.182	0.003	0.198	0.225	0.158	0.003
	Cz	0.391	0.427	0.121	0.005	0.332	0.368	0.160	0.005	0.330	0.367	0.136	0.005
	GK	-0.061	0.082	0.120	0.018	-0.049	0.073	0.147	0.015	-0.050	0.075	0.123	0.016
	SoS	0.219	0.252	0.146	0.004	0.214	0.245	0.183	0.004	0.214	0.245	0.158	0.004
	SS2	0.139	0.158	0.153	0.002	0.111	0.127	0.198	0.002	0.110	0.127	0.173	0.002
	FM	0.400	0.436	0.121	0.005	0.332	0.368	0.160	0.005	0.330	0.367	0.136	0.005

Summarizing results obtained

Table 19- IPA observed and adjusted values and the corresponding coefficients of variation (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=2; EM	IPA	Dataset – separation											
		Poor				Moderate				Good			
		Obs.	c.v.	Adj.	cv	Obs	cv	Adj	cv	Obs	cv	Adj	cv
Balanced	R	0.527	0.056	0.055	1.062	0.865	0.026	0.729	0.061	0.981	0.007	0.961	0.014
	GL	0.690	0.036	0.071	1.044	0.927	0.014	0.782	0.049	0.990	0.003	0.971	0.010
	J	0.441	0.076	0.039	1.060	0.762	0.045	0.644	0.080	0.962	0.013	0.943	0.020
	Cz	0.612	0.053	0.055	1.062	0.865	0.026	0.729	0.061	0.981	0.007	0.961	0.014
	GK	0.121	0.919	0.121	0.919	0.950	0.019	0.950	0.019	0.999	0.001	0.999	0.001
	SoS	0.232	0.330	0.040	1.047	0.748	0.051	0.664	0.077	0.962	0.013	0.949	0.018
	SS2	0.284	0.096	0.025	1.059	0.617	0.072	0.522	0.107	0.927	0.025	0.909	0.032
	FM	0.626	0.071	0.056	1.044	0.865	0.026	0.729	0.061	0.981	0.007	0.961	0.014
Unbalanced	R	0.575	0.091	0.097	1.133	0.886	0.021	0.765	0.051	0.981	0.008	0.962	0.016
	GL	0.729	0.058	0.120	1.138	0.940	0.011	0.811	0.042	0.991	0.004	0.971	0.012
	J	0.489	0.152	0.074	1.141	0.824	0.031	0.698	0.065	0.968	0.013	0.946	0.022
	Cz	0.654	0.101	0.097	1.133	0.903	0.017	0.765	0.051	0.984	0.007	0.962	0.016
	GK	0.239	1.083	0.240	1.080	0.966	0.012	0.966	0.012	0.999	0.001	0.999	0.001
	SoS	0.269	0.354	0.078	1.095	0.779	0.045	0.709	0.064	0.962	0.016	0.950	0.021
	SS2	0.327	0.203	0.050	1.153	0.702	0.053	0.595	0.086	0.938	0.025	0.917	0.034
	FM	0.660	0.109	0.099	1.130	0.904	0.017	0.765	0.051	0.984	0.007	0.962	0.016

Table 20 -IPA observed and adjusted values and the corresponding coefficients of variation (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=2; KM	IPA	Dataset–separation											
		Poor				Moderate				Good			
		Obs.	c.v.	Adj.	cv	Obs	cv	Adj	cv	Obs	cv	Adj	cv
Balanced	R	0.572	0.038	0.143	0.304	0.859	0.020	0.718	0.047	0.978	0.010	0.956	0.021
	GL	0.727	0.025	0.182	0.296	0.924	0.011	0.772	0.038	0.989	0.005	0.967	0.016
	J	0.401	0.052	0.101	0.314	0.753	0.034	0.630	0.061	0.958	0.020	0.936	0.031
	Cz	0.572	0.038	0.143	0.304	0.859	0.020	0.718	0.047	0.978	0.010	0.956	0.021
	GK	0.280	0.296	0.280	0.296	0.946	0.016	0.946	0.016	0.999	0.001	0.999	0.001
	SoS	0.327	0.075	0.103	0.317	0.738	0.039	0.651	0.059	0.957	0.020	0.943	0.027
	SS2	0.251	0.065	0.063	0.322	0.605	0.054	0.506	0.081	0.919	0.038	0.899	0.049
	FM	0.572	0.038	0.143	0.304	0.859	0.020	0.718	0.047	0.978	0.010	0.956	0.021
Unbalanced	R	0.553	0.042	0.105	0.433	0.855	0.029	0.706	0.073	0.974	0.009	0.947	0.018
	GL	0.712	0.027	0.134	0.420	0.922	0.016	0.760	0.060	0.987	0.004	0.959	0.014
	J	0.416	0.054	0.075	0.447	0.774	0.044	0.627	0.092	0.956	0.015	0.926	0.026
	Cz	0.587	0.038	0.105	0.433	0.872	0.025	0.706	0.073	0.977	0.008	0.947	0.018
	GK	0.211	0.418	0.211	0.418	0.944	0.023	0.944	0.023	0.998	0.001	0.998	0.001
	SoS	0.303	0.085	0.075	0.453	0.728	0.060	0.640	0.090	0.948	0.018	0.931	0.024
	SS2	0.263	0.068	0.047	0.460	0.633	0.072	0.513	0.119	0.916	0.029	0.887	0.039
	FM	0.588	0.038	0.105	0.432	0.873	0.025	0.706	0.072	0.977	0.008	0.947	0.018

Table 21 - IPA observed and adjusted values and the corresponding coefficients of variation (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=4; EM	IPA	Dataset – separation											
		Poor				Moderate				Good			
		Obs.	c.v.	Adj.	cv	Obs	cv	Adj	cv	Obs	cv	Adj	cv
Balanced	R	0.535	0.110	0.041	0.502	0.858	0.015	0.624	0.052	0.946	0.008	0.855	0.024
	GL	0.695	0.075	0.053	0.484	0.923	0.008	0.671	0.045	0.972	0.004	0.879	0.020
	J	0.213	0.073	0.025	0.503	0.561	0.050	0.487	0.069	0.805	0.031	0.772	0.038
	Cz	0.351	0.061	0.041	0.502	0.719	0.033	0.624	0.052	0.891	0.017	0.855	0.024
	GK	0.108	0.448	0.108	0.448	0.920	0.019	0.920	0.019	0.991	0.003	0.991	0.003
	SoS	0.235	0.065	0.031	0.507	0.651	0.042	0.569	0.060	0.859	0.023	0.827	0.029
	SS2	0.119	0.082	0.014	0.503	0.391	0.069	0.339	0.088	0.674	0.053	0.646	0.059
	FM	0.370	0.092	0.044	0.480	0.719	0.033	0.624	0.052	0.891	0.017	0.855	0.024
Unbalanced	R	0.565	0.076	0.133	0.437	0.919	0.008	0.820	0.022	0.949	0.008	0.887	0.018
	GL	0.721	0.050	0.169	0.425	0.958	0.004	0.855	0.018	0.974	0.004	0.910	0.014
	J	0.322	0.123	0.089	0.454	0.789	0.026	0.734	0.033	0.862	0.021	0.825	0.027
	Cz	0.485	0.095	0.133	0.437	0.882	0.015	0.820	0.022	0.926	0.011	0.887	0.018
	GK	0.296	0.420	0.297	0.420	0.982	0.004	0.982	0.004	0.993	0.002	0.993	0.002
	SoS	0.309	0.123	0.100	0.451	0.827	0.020	0.777	0.027	0.890	0.017	0.857	0.022
	SS2	0.192	0.145	0.053	0.467	0.653	0.043	0.607	0.050	0.758	0.037	0.726	0.043
	FM	0.496	0.103	0.138	0.435	0.882	0.015	0.820	0.022	0.926	0.011	0.887	0.018

Table 22 - IPA observed and adjusted values and the corresponding coefficients of variation (values are averaged over the 30 datasets and correspond to external validation of the clusters obtained).

K=4 KM	IPA	Dataset–separation											
		Poor				Moderate				Good			
		Obs.	c.v.	Adj.	cv	Obs	cv	Adj	cv	Obs	cv	Adj	cv
Balanced	R	0.635	0.006	0.029	0.327	0.814	0.044	0.518	0.155	0.929	0.032	0.812	0.086
	GL	0.777	0.003	0.035	0.325	0.897	0.025	0.569	0.137	0.963	0.018	0.842	0.074
	J	0.157	0.030	0.017	0.331	0.476	0.126	0.385	0.194	0.758	0.089	0.716	0.115
	Cz	0.272	0.026	0.029	0.327	0.643	0.086	0.518	0.155	0.860	0.057	0.812	0.086
	GK	0.074	0.321	0.074	0.321	0.847	0.076	0.847	0.076	0.981	0.023	0.981	0.023
	SoS	0.206	0.029	0.022	0.330	0.564	0.113	0.461	0.177	0.820	0.074	0.778	0.099
	SS2	0.085	0.033	0.009	0.333	0.314	0.164	0.255	0.231	0.614	0.125	0.581	0.147
	FM	0.272	0.026	0.029	0.327	0.644	0.085	0.519	0.153	0.861	0.055	0.813	0.084
Unbalanced	R	0.606	0.013	0.067	0.200	0.830	0.068	0.608	0.225	0.902	0.018	0.778	0.050
	GL	0.754	0.008	0.084	0.196	0.906	0.037	0.660	0.196	0.949	0.010	0.818	0.042
	J	0.204	0.035	0.041	0.205	0.583	0.215	0.489	0.300	0.741	0.055	0.678	0.072
	Cz	0.339	0.029	0.067	0.200	0.729	0.137	0.608	0.225	0.851	0.032	0.778	0.050
	GK	0.162	0.193	0.162	0.193	0.893	0.082	0.893	0.082	0.975	0.010	0.975	0.010
	SoS	0.247	0.034	0.052	0.205	0.646	0.176	0.550	0.257	0.790	0.044	0.731	0.060
	SS2	0.114	0.039	0.023	0.209	0.422	0.301	0.357	0.383	0.590	0.085	0.540	0.102
	FM	0.343	0.029	0.069	0.200	0.732	0.133	0.611	0.220	0.852	0.031	0.780	0.049

Table 23- Mean and Variation coefficient of observed indices (averaged over the 30 datasets)

	EM-STAB-2K-BAL					
Dataset	Poor		Moderated		Good	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.717	0.203	0.993	0.005	0.994	0.007
RR	0.666	0.249	0.497	0.006	0.496	0.007
NMI1	0.185	1.114	0.973	0.020	0.973	0.025
NMI2	0.129	1.331	0.973	0.020	0.973	0.025
NMI3	0.121	1.415	0.973	0.020	0.973	0.025
1-NVI	0.912	0.041	0.994	0.004	0.994	0.005
GL	0.827	0.121	0.997	0.003	0.997	0.004
J	0.699	0.225	0.987	0.011	0.987	0.014
C	0.813	0.136	0.993	0.005	0.993	0.007
GK	0.145	3.496	1.000	0.000	1.000	0.001
SoS	0.203	1.055	0.987	0.011	0.987	0.014
SS2	0.560	0.342	0.975	0.021	0.975	0.027
FM	0.823	0.125	0.993	0.005	0.993	0.007

Table 24 Mean and Variation coefficient of adjusted indices (averaged over the 30 datasets)

	EM-STAB-2K-BAL					
Dataset	Poor		Moderated		Good	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.111	2.033	0.987	0.011	0.987	0.014
RR	0.049	2.076	0.329	0.010	0.329	0.014
NMI1	0.178	1.178	0.973	0.020	0.973	0.025
NMI2	0.125	1.385	0.973	0.020	0.973	0.025
NMI3	0.117	1.464	0.973	0.020	0.973	0.025
1-NVI	0.117	1.464	0.973	0.020	0.973	0.025
GL	0.122	1.976	0.990	0.008	0.990	0.011
J	0.099	2.114	0.981	0.016	0.981	0.021
C	0.111	2.033	0.987	0.011	0.987	0.014
GK	0.204	2.262	1.000	0.000	1.000	0.001
SoS	0.107	2.033	0.983	0.014	0.983	0.019
SS2	0.084	2.244	0.968	0.026	0.969	0.034
FM	0.113	2.008	0.987	0.011	0.987	0.014

Table 25- Mean and Variation coefficient of observed indices (averaged over the 30 datasets)

	KM-STAB-2K-BAL					
Dataset	Poor		Moderated		Good	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.859	0.120	0.997	0.004	0.996	0.005
RR	0.431	0.119	0.499	0.006	0.498	0.005
NMI1	0.660	0.314	0.988	0.017	0.982	0.017
NMI2	0.659	0.315	0.987	0.017	0.982	0.017
NMI3	0.659	0.315	0.987	0.017	0.982	0.017
1-NVI	0.924	0.050	0.997	0.004	0.996	0.004
GL	0.921	0.071	0.999	0.002	0.998	0.002
J	0.767	0.189	0.994	0.008	0.992	0.009
C	0.860	0.119	0.997	0.004	0.996	0.005
GK	0.901	0.195	1.000	0.000	1.000	0.000
SoS	0.749	0.219	0.994	0.008	0.992	0.009
SS2	0.642	0.280	0.989	0.016	0.984	0.017
FM	0.860	0.119	0.997	0.004	0.996	0.005

Table 26- Mean and Variation coefficient of adjusted indices (averaged over the 30 datasets)

	KM-STAB-2K-BAL					
Dataset	Poor		Moderated		Good	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.718	0.287	0.994	0.008	0.992	0.009
RR	0.240	0.287	0.331	0.009	0.330	0.009
NMI1	0.660	0.315	0.987	0.017	0.982	0.017
NMI2	0.658	0.315	0.987	0.017	0.982	0.017
NMI3	0.658	0.315	0.987	0.017	0.982	0.017
1-NVI	0.658	0.315	0.987	0.017	0.982	0.017
GL	0.762	0.256	0.996	0.006	0.994	0.007
J	0.649	0.336	0.991	0.012	0.988	0.013
C	0.718	0.287	0.994	0.008	0.992	0.009
GK	0.901	0.195	1.000	0.000	1.000	0.000
SoS	0.665	0.329	0.992	0.011	0.989	0.012
SS2	0.551	0.407	0.986	0.020	0.980	0.022
FM	0.718	0.287	0.994	0.008	0.992	0.009

Table 27- Mean and Variation coefficient of observed indices(averaged over the 30 datasets)

	EM-STAB-3K-BAL					
Dataset	Poor		Mod		Good	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.628	0.215	0.827	0.104	0.990	0.005
RR	0.497	0.375	0.419	0.268	0.335	0.013
NMI1	0.138	**1.031**	0.695	0.120	0.964	0.016
NMI2	0.107	**1.063**	0.636	0.143	0.964	0.016
NMI3	0.103	**1.084**	0.632	0.148	0.964	0.016
1-NVI	0.861	0.059	0.910	0.035	0.988	0.005
GL	0.763	0.131	0.903	0.059	0.995	0.002
J	0.562	0.298	0.702	0.209	0.972	0.014
C	0.706	0.189	0.816	0.129	0.986	0.007
GK	0.276	**1.264**	0.915	0.091	1.000	0.000
SoS	0.267	0.507	0.689	0.192	0.979	0.010
SS2	0.412	0.433	0.561	0.309	0.946	0.026
FM	0.714	0.183	0.821	0.122	0.986	0.007

Table 28- Mean and Variation coefficient of adjusted indices (averaged over the 30 datasets)

	EM-STAB-3K-BAL					
Dataset	Poor		Mod		Good	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.123	**1.280**	0.651	**0.252**	0.979	**0.011**
RR	0.050	**1.266**	0.222	**0.320**	0.248	**0.014**
NMI1	0.132	**1.091**	0.694	**0.120**	0.964	**0.016**
NMI2	0.103	**1.116**	0.635	**0.143**	0.964	**0.016**
NMI3	0.098	**1.135**	0.631	**0.149**	0.964	**0.016**
1-NVI	0.098	**1.135**	0.631	**0.149**	0.964	**0.016**
GL	0.143	**1.244**	0.706	**0.212**	0.983	**0.008**
J	0.101	**1.318**	0.566	**0.323**	0.965	**0.017**
C	0.123	**1.280**	0.651	**0.252**	0.979	**0.011**
GK	0.292	**1.145**	0.915	**0.091**	1.000	**0.000**
SoS	0.110	**1.293**	0.592	**0.292**	0.973	**0.013**
SS2	0.076	**1.375**	0.457	**0.414**	0.940	**0.030**
FM	0.125	**1.267**	0.657	**0.241**	0.979	**0.011**

Table 29- Mean and Variation coefficient of observed indices (averaged over the 30 datasets)

	KM-STAB-3K-BAL					
Dataset	Poor		Mod		Good	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.852	0.090	0.884	0.100	0.994	0.004
RR	0.266	0.133	0.285	0.146	0.335	0.009
NMI1	0.661	0.213	0.730	0.225	0.975	0.014
NMI2	0.657	0.216	0.727	0.228	0.975	0.014
NMI3	0.657	0.216	0.727	0.228	0.975	0.014
1-NVI	0.889	0.051	0.912	0.058	0.992	0.005
GL	0.918	0.050	0.936	0.056	0.997	0.002
J	0.658	0.221	0.731	0.233	0.981	0.012
C	0.784	0.141	0.832	0.150	0.991	0.006
GK	0.901	0.126	0.924	0.122	1.000	0.000
SoS	0.703	0.200	0.767	0.212	0.986	0.009
SS2	0.507	0.320	0.602	0.331	0.964	0.023
FM	0.784	0.141	0.832	0.150	0.991	0.006

Table 30- Mean and Variation coefficient of adjusted indices (averaged over the 30 datasets)

	KM-STAB-3K-BAL					
Dataset	A		Poor		Mod	
IA	Mean	V. C.	Mean	V. C.	Mean	V. C.
Rand	0.672	0.252	0.744	0.259	0.986	0.009
RR	0.170	0.246	0.189	0.254	0.249	0.010
NMI1	0.660	0.214	0.729	0.225	0.975	0.014
NMI2	0.657	0.217	0.726	0.228	0.975	0.014
NMI3	0.657	0.217	0.726	0.228	0.975	0.014
1-NVI	0.657	0.217	0.726	0.228	0.975	0.014
GL	0.719	0.219	0.781	0.225	0.989	0.007
J	0.569	0.323	0.660	0.328	0.977	0.015
C	0.672	0.252	0.743	0.259	0.986	0.009
GK	0.901	0.126	0.924	0.122	1.000	0.000
SoS	0.616	0.294	0.699	0.301	0.982	0.012
SS2	0.444	0.416	0.550	0.412	0.959	0.026
FM	0.672	0.252	0.744	0.259	0.986	0.009

APPENDIX D – LIST OF CONFERENCES PRESENTATIONS

Table 31- List of conferences presentations

Conference	Year	
SPE	2010	**Title:** Limiares de concordância entre duas partições **Abstract**: Neste trabalho determina-se a significância da concordância entre duas partições, medida através dos índices de Rand e NMI - Normalized Mutual Information. Para o efeito são determinados os valores destes índices associados a tabelas de classificação cruzada, geradas sob a hipótese de não concordância (independência) restrita. A análise de dados é efectuada sobre uma base de dados com estrutura de agrupamento conhecida à qual se associam partições alternativas.
ICC	2011	**Title** Measuring the agreement between partitions: the use of thresholds values **Abstract:** The property of stability is often considered when evaluating the quality of a clustering solution. In particular, the solution's reproducibility in diverse data sets drawn from the same source may be considered as an indicator of stability. In order to measure stability one can use indices of agreement (IA) between the alternative partitions obtained from the diverse data sets. In fact, there are countless IA which can be used for this end. However, there has been few investment concerning the determination of IA thresholds values which can help deciding how much agreement is enough to derive stability (the well known Hubert and Arabie adjusted Rand index is an important exception). In the present work, we propose using simulated IA values, corresponding to cross-classification tables generated under the hypothesis of restricted independence (table with fixed marginal totals), to obtain IA thresholds. The Rand index, Mutual Information and the Variation of Information are used as IA examples. The R software is used to implement the proposed approach. Four simulated data sets (Gaussian, mixture model based, with different degrees of separation) are used to obtain experimental results. The stability of alternative clustering solutions provided by different clustering algorithms (K-Means and EM type) is evaluated and discussed using a cross-validation approach. In addition, the agreement with the real partition is also discussed.

SPE	2011	**Title** Índices de informação mútua na avaliação de estabilidade de agrupamentos **Abstract:** Neste trabalho avalia-se o desempenho de diversos índices de informação mútua no papel de indicadores da estabilidade de partições. Nesta avaliação são determinadas estimativas dos valores dos índices sob hipótese de independência restrita. A análise de dados é efectuada sobre quatro bases de dados com estruturas de agrupamento conhecidas, às quais se associam partições alternativas.
IDA	2011	**Title** Indices of agreement between partitions: a comparative approach **Abstract:** Indices of agreement (IA) between partitions are often used for external evaluation of clustering results as well as for the evaluation of clustering stability, i.e. the partition's reproducibility in diverse data sets drawn from the same source, e.g. (Gordon 1999). There are countless IA which can be used for this end. However, there has been little investment concerning its comparison taking into account the determination of the indices threshold values. In the present work, we propose using simulated IA values, corresponding to crossc1assitication tables generated under the hypothesis of restricted independence to obtain IA thresholds (Patefield 1981) and then compare the performance of several IA indices. The indices of Rand (Rand 1971), Russel and Rao (Russel and Rao 1940), Normalized Mutual Information and Variation of Information (Meila 2007) are used as IA examples, their interpretation being considered. Four simulated data sets (Gaussian, mixture model based, with different degrees of separation) are used to obtain the experimental results (Maitra and Melnykov 2010). The stability of alternative clustering solutions provided by different clustering algorithms (K-Means and EM type) is evaluated and discussed using a cross-validation approach (Cardoso, Carvalho and Faceli 2009). In addition, the agreement with the real partition is also discussed.
JOCLAD	2012	**Title** Índices de concordância pareada na avaliação de estabilidade de agrupamentos **Abstract:** Neste trabalho estuda-se o desempenho de diversos índices de concordância pareada como indicadores da estabilidade de partições. Para o efeito, são determinadas estimativas dos valores dos índices sob a hipótese de independência restrita. A análise de dados é efectuada sobre quatro bases de dados com estruturas de agrupamento conhecidas, às quais se associam partições alternativas.

SMTDA	2014	**Title** The influence of classes entropy and overlap on Random Forests performance **Abstract**: In order to evaluate the impact of class‟overlap and entropy on the performance of Random Forests we conduct some experiments base on synthetic data-360 data sets are generated. We set up the scenarios for our experiments by considering different classification problems with 2, 3 and 4 classes, diverse degrees of classes „entropy and overlap. According to the obtained results, the average performance of random forests significantly decreases with the increase of the degree of classes‟ overlap and this impact surpassed the impact of the classes entropy. Statistical analysis conducted yield additional insights referring to diverse measures of classification performance.
JOCLAD	2015	**Title** Distribuições de índices de concordância entre agrupamentos **Abstract:** No presente trabalho estuda-se a distribuição empírica de diversos índices de concordância pareada entre duas partições, sob a hipótese de independência restrita (H_o). A distribuição de cada índice é obtida a partir de um processo de simulação, sendo os seus valores resultantes da geração de tabelas de contingência, sob H_o. Estas tabelas correspondem à avaliação externa de agrupamentos que se realizam em diferentes cenários, controlando os números de grupos, dimensões relativas e graus de sobreposição dos grupos. Os cenários de agrupamento correspondem a bases de dados simulados (misturas de Normais).

AGRESTI, A., WACKERLY, D. & BOYETT, J. M. 1979. Exact conditional tests for cross-classifications: approximation of attained significance levels. *Psychometrika,* 44**,** 75-83.

ALBATINEH, A. N. & NIEWIADOMSKA-BUGAJ, M. 2011. Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Adv. Data Anal. Classification,* 5**,** 179-200.

ALIZADEH, H., MIMAEI-BIDGOLI, B. & PARVIN, H. 2014. To improve the quality of cluster ensembles by selecting a subset of base clusters. *Journal of Experimental and Theoretical Artificial Intelligence,* 26**,** 127-150.

AMORIM, M. J. & CARDOSO, M. G. M. S. Paired Indices for clustering Evaluation. Correction for Agreement by Chance. *In:* THE, I. P. O., ed. 16 th International Conference on Enterprise Information Systems, 2014 Lisboa, Portugal, 27-30 Abril. 164-170.

AMORIM, M. J. & CARDOSO, M. G. M. S. 2015a. Clustering stability and ground truth: numerical experiments. *International Journal of Artificial Intelligence and Knowledge Discovery,* to appear.

AMORIM, M. J. & CARDOSO, M. G. M. S. Clustering stability and ground truth: numerical experiments. 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - , 2015b Lisboa. Science and Technology Publications, 259-264.

AMORIM, M. J. & CARDOSO, M. G. M. S. 2015c. Comparing clustering solutions: The use of adjusted paired indices. *Intelligent Data Analysis,* 19**,** 1275–1296.

AMORIM, M. J. P. C. & CARDOSO, M. G. M. S. Clustering cross-validation and mutual information indices. *In:* ANA COLUBI, K. F., GIL GONZALEZ-RODRIGUEZ AND ERRICOS JOHN KONTOGHIORGHES, ed. 20th International Conference on Computational Statistics (COMPSTAT 2012), 2012 Limassol, Cyprus. The International Statistical Institute/International Association for Statistical Computing, 39-52.

BACHE, K. & LICHMAN, M. 2013. *UCI Machine Learning Repository* [Online]. Irvine, CA: University of California, School of Information and Computer Science. . Available: http://archive.ics.uci.edu/ml.

BEN-DAVID, S. & LUXBURG, U. V. Relating clustering stability to properties of cluster boundaries. *In:* SERVEDIO, R. & ZHANG, T., eds. 21st Annual Conference on Learning Theory (COLT), July, 9-12 2008 Berlin. Springer, 379-390.

BEN-DAVID, S., SIMON, H. U. & P´AL, D. A. A Sober Look at Clustering Stability. Conference on Computational Learning Theory, 2006. 5–19.

BEN-HUR, A., ELISSEEFF, A. & GUYON, I. 2002. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing,* 7**,** 6-17.

BIERNACKI, C., CELEUX, G., GOVAERT, G. & LANGROGNET, F. 2006. Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis,* 51**,** 587-600.

CARDOSO, M. G., FACELI, K. & DE CARVALHO, A. C. 2010. Evaluation of Clustering Results: The Trade-off Bias-Variability. *Classification as a Tool for Research.* Springer.

CARDOSO, M. G. M. S., FACELI, K. & CARVALHO, A. P. D. L. F. D. Evaluation of clustering results: the trade-off bias-variability. *In:* LOCARECK-JUNGE, H. & WEIHS, C., eds. IFCS 2009 – 11th Conference of the International Federation of Classification Societies, 2009 Dresden. 201-208.

CHENG, R. & MILLIGAN, G. W. 1996. Measuring the Influence of Individual. Data Points in a Cluster Analysis. *Journal of Classification,* 13**,** 315-335.

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 51**,** 821-828.

CZEKANOWSKI, J. 1932. "Coefficient of racial likeness" and "durchschnittliche Differenz". *Anthropologischer Anzeiger,* 14**,** 227-249.

DOLNICAR, S. & LEISCH, F. 2010. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Market Lett,* 21**,** 83–101.

DUDOIT, S. & FRIDLYAND, J. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology,* 3**,** 1-21.

FANG, Y. & WANG, J. 2012. Selection of the number of clusters via the bootstrap method. *Computational Statistics and Data Analysis,* 56**,** 468–477.

FOWLKES, E. B. & MALLOWS, C. L. 1983. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association,* 78**,** 553-569.

FRED, A. & JAIN, A. K. 2003. Robust Data Clustering. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* USA.

GOODMAN, L. A. & KRUSKAL, W. H. 1954. Measures of Association for Cross Classifications. *Journal of the American Statistical Associations,* 49.

GOWER, J. C. & LEGENDRE, P. 1986. Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification,* 3.

HALTON, J. H. 1969. A Rigorous Derivation of the Exact Contingency Formula. *Proc. Camb. Phil. Soc,* 65

527-530.

HARTIGAN, J. A. 1975. *Clustering algorithms.*

HENNIG, C. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis,* 52**,** 258-271.

HENNIG, C. & LIAO, T. F. 2013. How to find an appropriate clustering for mixed type variables with application to socio-economic stratification. *Appl. Statist.,* 62**,** 309–369.

HUBERT, L. & ARABIE, P. 1985. Comparing partitions. *Journal of Classification,* 2**,** 193-218.

JACCARD 1908. Nouvelles recerches sur la distribuition florale. *Bulletin de la Societé Vaudoise de Sciences Naturells,* 44**,** 223-370.

KREY, S., BRATOB, S., LIGGESA, U., GÖTZEB, J. & WEIHSA, C. 2015. Clustering of electrical transmission systems based on network topology and stability. *Journal of Statistical Computation and Simulation,,* 85 47–61.

KREY, S., LIGGES, U. & LEISCH, F. 2014. Music and timbre segmentation by recursive constrained K-means clustering. *Computational Statistics* 29**,** 37-50.

LANGE, T., ROTH, V., BRAUN, M. L. & BUHMANN, J. M. 2004. Stability-Based Validation of Clustering Solutions. *Neural Computation,* 16**,** 1299–1323.

LUXBURG, U. V. 2009. Clustering Stability: An Overview. *Machine Learning,* 2**,** 235-274.

MAITRA, R. & MELNYKOV, V. 2010. Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics,* 19 354-376.

MARATEB, H. R., MANSOURIAN, M., ADIBI , P. & FARINA, D. 2014. Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies. *Journal of Research in Medical Sciences,* 19**,** 47-56.

MEILÃ, M. 2007. Comparing Clusterings - an information based distance. *Journal of Multivariate Analysis,* 98**,** 873-895.

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,* 52**,** 91-118.

MÜLLER, H. & HAMM, U. 2014. Stability of market segmentation with cluster analysis – A methodological approach. *Food Quality and Preference,* 34**,** 70-78.

PASCUAL, D., PLA , F. & SÁNCHEZ, J. S. 2010. Cluster validation using information stability measures. *Pattern Recognition Letters,* 31**,** 454-461.

PATEFIELD, W. M. 1981. Algorithm As159: An Efficient Method of Generating Random R * C Tables Given Row. *Roayal Statistical Society,* Series c, Applied Statistics**,** 91-97.

RAND, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association,* 66**,** 846-850.

RAVERA, O. 2001. A comparison between diversity, similarity and biotic indices applied to the macroinvertebrate community of a small stream: the Ravella river (Como Province, Northern Italy). *Aquatic Ecology,* 35**,** 97–107.

ROTH, V., BRAUN, M. L., LANGE, T. & BUHMANN, J. M. 2002. Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data. *J.R. Dorronsoro (Ed.): ICANN 2002, LNCS 2415***,** 607–612.

SHAMIR, O. & TISHBY, N. Cluster stability for finite samples. *In:* J. C. PLATT, D. K., Y. SINGER & S. ROWEIS ed. Advances in neural information processing systems 2008. Cambridge:MIT Press., 1297–1304.

SHAMIR, O. & TISHBY, N. 2010. Stability and model selection in k-means clustering. *Mach Learn,* 80**,** 213-244.

SOKAL, R. R. & SNEATH, P. H. 1963. *Principles of Numerical Taxonomy,* San Francisco CA:Freeman.

STREHL, A. & GOHOSH, J. 2002. Cluster Ensembles- a Knowledge Reuse Framework for Combinig Partitions. *Journal of Machine Learning Research,* 3**,** 583-617.

VINH, N. X., EPPS, J. & BAILEY, J. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research,* 9999**,** 2837-2854.

WANG, J. 2010. Consistent selection of the number of clusters via crossvalidation. *Biometrika,* 97**,** 893–904.

WU, H.-M. 2011. On biological validity indices for soft clustering algorithms for gene expression data. *Computational Statistics and Data Analysis,* 55**,** 1969-1979.