Une Méthodologie pour la Comparaison de Partitions

Genane Youness, Gilbert Saporta

▶ **To cite this version:**

Genane Youness, Gilbert Saporta. Une Méthodologie pour la Comparaison de Partitions. Revue de Statistique Appliquée, 2004, 52, pp.97-120. hal-01124920

HAL Id: hal-01124920
https://hal.science/hal-01124920

Submitted on 12 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

REVUE DE STATISTIQUE APPLIQUÉE

GENANE YOUNESS
GILBERT SAPORTA

Une méthodologie pour la comparaison de partitions

Revue de statistique appliquée, tome 52, n° 1 (2004), p. 97-120
<http://www.numdam.org/item?id=RSA_2004__52_1_97_0>

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 2004, **LII** (1), 97-120

UNE MÉTHODOLOGIE POUR LA COMPARAISON DE PARTITIONS

Genane YOUNESS*, Gilbert SAPORTA**

* *Université Pierre et Marie Curie et ISAE-CNAM , BP 11 4661, Beyrouth, Liban, genane99@hotmail.com*

** *Chaire de Statistique Appliquée-CEDRIC, CNAM, 292 rue Saint-Martin, 75003 Paris, France, saporta@cnam.fr*

RÉSUMÉ

Nous proposons une méthodologie pour comparer des partitions d'un même ensemble de données. Nous présentons tout d'abord quelques mesures de comparaison de deux classifications d'un même ensemble de données : l'indice de Rand, sous sa forme brute ou corrigée, ainsi que sa version asymétrique, puis deux autres indices : le premier est inspiré du test de Mac Nemar et le second de l'indice de Jaccard. On présente les écritures logiques et relationnelles de ces indices ainsi que leurs distributions d'échantillonnage sous une hypothèse nulle d'absence de liaison. Pour étudier la stabilité des classes on utilise ensuite un modèle particulier de mélanges de distributions, les profils latents qui sert à simuler des données que l'on classe selon deux groupes de variables.

Mots-clés : *Classes latentes, K-means, Indices d'associations, Tests statistiques, Partitions*

ABSTRACT

We propose a methodology for comparing two partitions of the same data set. The study begins by presenting some measures of comparison: the Rand's measure of association, the asymmetric, and the corrected one. We also study two other indices: the first is based on an adaptation of Mac Nemar's test, the second being Jaccard's index. We present the logic form based on pair comparisons for this indexes as well as their sampling distribution under the assumption of independence. Stability of classes is studied by simulating data coming from a latent profile model which is a particular modal of a mixture distribution and we partition them according to 2 groups of variables.

Keywords : *Latent class, K-means, Indices of association, Statistical test, Partitions*

1. Introduction

Quand on dispose de deux partitions effectuées sur les mêmes individus, par exemple avec deux jeux de variables ou bien avec deux algorithmes, la question se pose naturellement de savoir si ces deux partitions sont en accord ou bien si elles diffèrent significativement, en un sens à préciser. Une manière d'aborder ce problème consiste à calculer un indice de concordance entre partitions et à définir une valeur critique

au-dessus ou en deçà de laquelle on conclura que les deux partitions sont ou non concordantes. À l'indice bien connu de Rand, nous proposons d'ajouter deux indices inspirés du test de Mac Nemar et de l'indice de Jaccard. Une version asymétrique de Rand [CHAV 01] sera utilisée pour la comparaison de partitions emboîtées, avec des nombres différents de classes.

Il faut alors connaître, au moins approximativement, la distribution de probabilités de ces indices. Mais sous quelle hypothèse ? Cette question ne semble curieusement pas avoir été traitée dans la littérature, en tous cas pas sous des hypothèses réalistes [SAP 97], [SAP 02]. En effet, les rares travaux connus et récents [IDR 00], concernent la distribution de l'indice de Rand sous l'hypothèse d'indépendance. Or cette hypothèse n'est évidemment pas pertinente pour la question posée, car la non-indépendance ne signifie pas nécessairement une forte concordance.

La difficulté est de conceptualiser une hypothèse nulle d'identité de deux partitions. On se trouve dans une situation voisine de celle où on voudrait tester que deux variables numériques sont identiques : or si $\rho = 1$, on sait bien que $r = 1$ et on n'a donc pas de test utile de l'hypothèse nulle qui se trouve rejetée dès que $r < 1$.

Notre approche consiste à étudier la distribution des indices d'associations en engendrant par simulation des partitions qui devraient être proches car issues d'un même modèle sous-jacent : nous utiliserons pour cela un modèle de classes latentes régissant la distribution de k variables. Nous partageons ensuite arbitrairement les k variables en deux groupes et nous comparons les partitions engendrées par une méthode classique de nuées dynamiques sur chacun de ces groupes de variables.

2. Indices de Comparaison de partitions

Comparer deux partitions revient d'une certaine façon à comparer deux variables qualitatives et on pourrait penser à utiliser des indices de liaison bien connus comme le chi-2 et ses dérivés (phi-2 de Pearson, V de Cramer ou T^2 de Tschuprow). Ces indices qui mesurent l'écart à l'indépendance sont en fait moins bien adaptés à la mesure de la liaison que ceux fondés sur les concordances entre paires d'unités statistiques, et c'est pour cela que nous ne les utiliserons pas.

2.1. Notations

$\mathcal{P}_1$ et $\mathcal{P}_2$ sont deux partitions des mêmes individus (ou deux variables qualitatives).

N désigne le tableau de contingence associé, K_1, K_2 les tableaux disjonctifs associés à $\mathcal{P}_1$ et $\mathcal{P}_2$; On a : $\mathbf{N} = \mathbf{K_1'K_2}$.

Chaque partition $\mathcal{P}_k$ sera représentée par un tableau relationnel C^k dans l'espace des individus, de dimension $n \times n$, dont le terme général $c^k_{ii'}$ est défini par :

$c^k_{ii'} = 1$ si i et i' sont deux individus dans la même classe de la partition $\mathcal{P}_k$
$= 0$ sinon

On a $C_1 = K_1K_1'$

On associe au tableau C son tableau complémentaire, notée $\overline{C}$ dont le terme général est défini par :

$$\overline{c}_{ii'}^{k} = 1 - c_{ii'}^{k} \begin{array}{ll} = 0 & \text{si } i \text{ et } i' \text{ sont deux individus dans la même classe de la partition } \mathcal{P}_k \\ = 1 & \text{sinon} \end{array}$$

Nous posons :

n = nombre d'individus

p = nombre de classes de la partition $\mathcal{P}_1$

q = nombre de classes de la partition $\mathcal{P}_2$

Lorsque l'on croise deux partitions, on va s'intéresser aux paires d'individus qui restent ou ne restent pas dans les mêmes classes. On a $\binom{n}{2}$ paires d'individus représentés par les 4 types suivants :

a : le nombre de paires dans une même classe de $\mathcal{P}_1$ et dans une même classe de $\mathcal{P}_2$

b : le nombre de paires séparées dans $\mathcal{P}_1$ et séparées dans $\mathcal{P}_2$

c : le nombre de paires séparées dans $\mathcal{P}_1$ et dans une même classe de $\mathcal{P}_2$

d : le nombre de paires dans une même classe de $\mathcal{P}_1$ et séparées dans $\mathcal{P}_2$

On notera également $A = a + b$ (nombre total d'accords) et $D = c + d$ (nombre total de désaccords)

On peut aussi, au lieu de considérer les $\binom{n}{2}$ paires (i, i') considérer les n^2 paires (i, i') (où (i, i') est distinguée de (i, i) et où l'on comptabilise les n paires (i, i)). Si a', b', c', d' désignent les équivalents de a, b, c, d, on a alors :

$$a' = 2a + n\,; \quad b' = 2b\,; \quad c' = 2c\,; \quad d' = 2d$$

2.2. Formules de linéarisation

Le tableau de contingence croisant $\mathcal{P}_1$ et $\mathcal{P}_2$ est de dimension $p \times q$. Il est caractérisé par son terme général :

$$n_{uv} = \text{l'effectif de la case } (u, v)$$

il est lié aux termes généraux des tableaux disjonctifs associées à $\mathcal{P}_1$ et $\mathcal{P}_2$ par la formule suivante :

$$n_{uv} = \sum_{i=1}^{n} k_{iu} k_{iv}$$

Les tableaux relationnels $\mathcal{C}^1$ et $\mathcal{C}^2$ fournissent la même information que le tableau de contingence croisant les partitions $\mathcal{P}_1$ et $\mathcal{P}_2$.

Les formules de passage contingences-paires ont été proposées et démontrées par Kendall [KEN 61] et Marcotorchino [MAR 84], rappelons celles qu'on va utiliser :

$$\sum_i \sum_{i'} c^1_{ii'} = \sum_u n^2_{u.}$$
$$\sum_i \sum_{i'} c^2_{ii'} = \sum_v n^2_{.v}$$

$$a' = 2a + n = \sum_i \sum_{i'} c^1_{ii'} c^2_{ii'} = \sum_u \sum_v n^2_{uv}$$
$$b' = 2b = \sum_i \sum_{i'} \bar{c}^1_{ii'} \bar{c}^2_{ii'} = \sum_i \sum_{i'} (1 - c^1_{ii'})(1 - c^2_{ii'})$$
$$= n^2 + \sum_u \sum_v n^2_{uv} - \sum_u n^2_{u.} - \sum_v n^2_{.v}$$
$$c' = 2c = \sum_i \sum_{i'} \bar{c}^1_{ii'} c^2_{ii'} = \sum_i \sum_{i'} (1 - c^1_{ii'}) c^2_{ii'}$$
$$= \sum_v n^2_{.v} - \sum_u \sum_v n^2_{uv}$$
$$d' = 2d = \sum_i \sum_{i'} c^1_{ii'} \bar{c}^2_{ii'} = \sum_u n^2_{u.} - \sum_u \sum_v n^2_{uv}$$

2.3. Indice de Rand symétrique

Dans le but de comparer deux partitions de taille p et q, l'indice d'accord le plus utilisé est l'indice de Rand. Cet indice brut de Rand (semblable au taux de Kendall pour les variables ordinales) est le pourcentage global de paires en accord :

$$R = \frac{A}{\binom{n}{2}}$$

On peut montrer que :

$$A = \binom{n}{2} + \sum_u \sum_v n^2_{uv} - \frac{1}{2}\left[\sum_u n^2_{u.} + \sum_v n^2_{.v}\right]$$

L'indice de Rand écrit sous sa forme contingentielle selon Marcotorchino [MAR 91] où on considère toutes les paires, y compris celles identiques est :

$$R' = \frac{2\sum_u \sum_v n^2_{uv} - \sum_u n^2_{u.} - \sum_v n^2_{.v} + n^2}{n^2}$$

Il prend ses valeurs entre 0 et 1; il est égal à 1 lorsque les deux partitions sont identiques.

En utilisant les formules de linéarisation, EL Ayoubi (N.) [MAR 91] a montré que cette dernière version de R peut être écrite sous la forme relationnelle suivante :

$$R' = \frac{1}{n^2}\left[\sum_i \sum_{i'} c^1_{ii'} c^2_{ii'} + \sum_i \sum_{i'} \bar{c}^1_{ii'} \bar{c}^2_{ii'}\right]$$

C'est avec cette formulation relationnelle qu'Idrissi [IDR 00] a étudié la normalité asymptotique de R' sous l'hypothèse d'indépendance. Par exemple si les k classes (dans le cas où les deux partitions ont même nombre de classes $p = q = k$) sont équiprobables on trouve que suit $c^1_{ii'} c^2_{ii'} + c^{-1}_{ii'} c^{-2}_{ii'}$ une loi de Bernoulli de paramètre : $1 - \frac{2}{k} + \frac{2}{k^2}$, on en déduit :

$$E(R') = 1 - \frac{2}{k} + \frac{2}{k^2}$$

A. Idrissi affirme ensuite que le coefficient de Rand empirique entre deux variables qualitatives à k modalités équiprobables calculées sur n observations suit asymptotiquement une loi normale de variance :

$$V(R') = \frac{1}{n^2}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{k} + \frac{2}{k^2}\right)\left(\frac{2}{k} - \frac{2}{k^2}\right)$$

Cette expression de la variance suppose l'indépendance des $c_{ii'}$, ce qui est inexact en raison des contraintes de transitivité ($c_{ik} = c_{ii'} c_{i'k}$) et n'est vraie qu'approximativement pour k grand (il n'y a même pas normalité asymptotique pour des partitions en deux classes).

2.4. Indice de Rand corrigé selon Huber et Arabie

Pour deux partitions aléatoires, la valeur espérée de l'indice de Rand n'est pas nulle. L'indice de Rand ajusté proposé par [HUB 85] a pour forme générale :

$$\frac{\text{indice} - \text{indice espéré}}{\text{indice maximum} - \text{indice espéré}}$$

Cet indice qui peut être au plus égal à 1, prend donc la valeur 0 quand l'indice = indice espéré.

Avec une hypothèse de distribution hypergéométrique, on montre que l'indice de Rand corrigé est égal à :

$$RC = \frac{R - R_{\text{esp}}}{R_{\max} - R_{\text{esp}}} = \frac{n^2 \cdot \sum_{u,v} n_{uv}^2 - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}{\frac{1}{2} \cdot n^2 \cdot \left(\sum_u n_{u.}^2 + \sum_v n_{.v}^2 \right) - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}$$

L'indice maximum $R_{\max}$ étant égal à 1, tandis que l'indice espéré R_{esp} s'obtient en remplaçant dans l'expression de R n_{uv} par $\frac{n_{u.} \cdot n_{.v}}{n}$. On peut noter qu'on aurait obtenu le même coefficient RC, si on avait fait le calcul à partir de R'.

L'indice de Rand brut est souvent plus élevé que celui corrigé. Hubert et Arabie affirment que la correction augmente la sensibilité de cet indice.

L'espérance de l'indice corrigé est nulle lorsque les accords entre les deux partitions sont dus au hasard. Cependant cet indice corrigé peut prendre des valeurs négatives lorsque les partitions sont peu liées.

2.5. Indice de Rand dans sa version asymétrique

Dans le cas où on a deux partitions d'un même ensemble d'individus mais avec des nombres de classes inégaux, on utilisera l'indice de Rand asymétrique proposé par [CHAV 01].

Cet indice asymétrique évalue dans quelle mesure une partition $\mathcal{P}_1$ (souvent experte) est « plus fine » qu'une partition $\mathcal{P}_2$. Lorsque la partition experte est engendrée par une variable qualitative, on peut simplement vouloir qu'une classe de la partition obtenue contienne tous les objets d'une ou de plusieurs classes de la partition experte $\mathcal{P}_1$. $\mathcal{P}_1$ aura alors en général plus de classes que $\mathcal{P}_2$ et il semble plus naturel d'utiliser des critères de comparaison non symétrique.

On considère les deux partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ de n individus dont le nombre de classes de $\mathcal{P}_1$ est supérieur au nombre de classes de $\mathcal{P}_2$. $\mathcal{P}_1$ est plus fine que $\mathcal{P}_2$ si, lorsque deux éléments sont classés ensemble dans $\mathcal{P}_1$, ils le sont également dans $\mathcal{P}_2 : \forall u = 1, ..., p$, $\exists v$ tel que $\mathcal{P}_u^1 \subseteq \mathcal{P}_v^2, \mathcal{P}_u^1$ (respectivement $\mathcal{P}_v^2$) désignant la u$^{\text{ème}}$ (respectivement v$^{\text{ème}}$) classe de $\mathcal{P}_1$ (respectivement $\mathcal{P}_2$).

On cherche ainsi à mesurer l'inclusion de la partition $\mathcal{P}_1$ dans la partition $\mathcal{P}_2$.

Nous présentons une écriture simple où nous considérons toutes les paires y compris celles identiques.

Ce critère de Rand asymétrique, noté RA, est défini par :

$$RA(P_1, P_2) = 1 + \frac{\sum_{u,v} \binom{n_{uv}}{2} - \sum_{u,v} \binom{n_{u\cdot}}{2}}{\binom{n}{2}}$$

Ce critère prend ses valeurs dans $[0,1]$. Si $\forall u, \exists v$ tel que $\mathcal{P}_u^1 \subseteq \mathcal{P}_v^2$, alors $RA = 1$.

En considérant toutes les paires d'individus, y compris celles identiques on peut écrire cette version de la façon suivante :

$$RA'(P_1,P_2) = \frac{n^2 + \sum_{u,v} n_{uv}^2 - \sum_u n_{u.}^2}{n^2} = \frac{a'+b'+c'}{a'+b'+c'+d'}$$

Notons que dans le cas où les deux partitions ont même nombre de classes, l'indice de Rand asymétrique n'est pas égal à l'indice de Rand brut.

Comme nous avons vu, il est difficile de déterminer les cas où R et RA sont nuls.

Le critère de Rand asymétrique corrigé utilise aussi la normalisation $\frac{RA - RA_{\text{esp}}}{1 - RA_{\text{esp}}}$ et vaut donc :

$$RA_c = \frac{n^2 \cdot \sum_{u,v} n_{uv}^2 - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}{n^2 \cdot \sum_u n_{u.}^2 - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}$$

En notant :

$N_{uv} = n_{uv}$ = nombre d'individus qui sont dans les classes u de $\mathcal{P}_1$ et v de $\mathcal{P}_2$

$N_{u.} = n_{u.}$ = nombre d'individus qui sont dans les classes u de $\mathcal{P}_1$

$N_{u\overline{v}} = n_{u.} - n_{uv}$ = nombre d'individus qui sont dans la classe u de $\mathcal{P}_1$ et ne sont pas dans la classe v de $\mathcal{P}_2$

$N_{\overline{u}v} = n_{.v} - n_{uv}$ = nombre d'individus qui sont dans la classe v de $\mathcal{P}_2$ et ne sont pas dans la classe u de $\mathcal{P}_1$

on obtient l'écriture suivante :

$$RA'(P_1,P_2) = \frac{1}{n^2}\sum_u\sum_v N_{uv}^2 + \frac{1}{n^2}\sum_u\sum_v N_{u.}N_{\overline{u}v}$$

qui peut être réécrite par les formules de comparaison par paires :

$$RA'(P_1,P_2) = 1 - \frac{1}{n^2}\sum_i\sum_{i'} c_{ii'}^1 \overline{c}_{ii'}^2$$

Sous l'hypothèse d'indépendance et d'équiprobabilité on trouve que $E(RA') = 1 - \frac{1}{p} + \frac{1}{pq}$.

2.6. L'indice dérivé de Mac Nemar

Le test de Mac Nemar est un test non-paramétrique bien connu utilisé pour vérifier l'égalité de deux proportions dans des échantillons appariés (par exemple pourcentage d'individus favorables à une certaine opinion avant et après une campagne).

a	c
d	b

Si a désigne le nombre d'individus qui ont gardé la même opinion favorable, avant et après, b le nombre d'individus qui ont gardé la même opinion défavorable, c et d les effectifs de ceux qui ont changé d'avis, la statistique de test correspondant à l'hypothèse nulle selon laquelle les changements d'opinion dans un sens ou d'autre sont équiprobables est :

$$M_c = \frac{d-c}{\sqrt{c+d}}$$

M_c suit approximativement une loi normale $N(0,1)$ sous H_0.

En adaptant le test de Mc Nemar à l'ensemble des paires d'individus, on a une nouvelle façon de mesurer la concordance entre deux partitions, qui revient à se demander si les paires qui sont séparées le sont par hasard, donc on étudie le désaccord entre les paires d'individus. Cet indice est d'autant plus proche de zéro que les deux partitions ne diffèrent que «par hasard» : sous réserve que c et d ne soient pas trop grands, ce sera un indice d'accord.

On montre facilement que :

$$M_c = \frac{\sum_u n_{u.}^2 - \sum_v n_{.v}^2}{2\sqrt{\frac{1}{2}\left(\sum_u n_{u.}^2 + \sum_v n_{.v}^2\right) - \sum_u \sum_v n_{uv}^2}}$$

$$= \frac{\sum_u n_{u.}^2 - \sum_v n_{.v}^2}{\sqrt{2}\sqrt{\sum_u n_{u.}^2 + \sum_v n_{.v}^2 - 2\sum_u \sum_v n_{uv}^2}}$$

on en déduit la forme relationnelle de cet indice :

$$M_c = \frac{\sum_i \sum_{i'} c_{ii'}^1 - \sum_i \sum_{i'} c_{ii'}^2}{\sqrt{2\left(\sum_i \sum_{i'} c_{ii'}^1 \bar{c}_{ii'}^2 + \sum_i \sum_{i'} \bar{c}_{ii'}^1 c_{ii'}^2\right)}}$$

2.7. Indice de Jaccard

L'indice de Jaccard est un coefficient d'association connu pour étudier la similarité entre objets pour des données binaires de présence-absence.

Le tableau binaire suivant représente un exemple de présence-absence de deux individus i et i' quelconques à m critères différents :

	v_1	v_2	v_3	...	v_m
i	1	1	0		1
i'	0	1	0		0

On peut former alors le tableau suivant :

i \ i'	1	0
1	11(i,i')	10(i,i')
0	01(i,i')	00(i,i')

où $11(i,i')$ = nombre de critères ou propriétés que i et i' possèdent simultanément

$01(i,i')$ = nombre de propriétés que i ne possède pas mais que i' possède

$10(i,i')$ = nombre de propriétés que i' ne possède pas mais que i possède

$00(i,i')$ = nombre de propriétés que i et i' ne possèdent pas.

L'indice de Jaccard est :

$$J(i,i') = \frac{11(i,i')}{11(i,i') + 10(i,i') + 01(i,i')}$$

Cet indice varie de 0 à 1 et ne tient compte que des associations positives (présences simultanées).

Par analogie avec ce cas, on définit l'indice de Jaccard d'accord entre deux partitions par :

$$J = \frac{a}{a+c+d}$$

on trouve alors que l'indice s'écrit :

$$J = \frac{\sum_u \sum_v n_{uv}^2 - n}{\sum_u n_{u.}^2 + \sum_v n_{.v}^2 - \sum_u \sum_v n_{uv}^2 - n}$$

Par les formules de passage on trouve la forme relationnelle de cet indice en utilisant les formules de passage contingence-paires :

$$J = \frac{\sum_i \sum_{i'} c^1_{ii'} c^2_{ii'} - n}{\sum_i \sum_{i'} \overline{c}^1_{ii'} c^2_{ii'} + \sum_i \sum_{i'} c^1_{ii'} - n}$$

3. Méthodologie de simulation

3.1. Un modèle de classes latentes

Il faut maintenant définir ce que l'on entend par « deux partitions sont proches » : notre approche consiste à dire que les individus proviennent d'une même partition commune, dont les deux partitions observées en sont des réalisations bruitées.

Le modèle de classes latentes est bien adapté à cette problématique pour engendrer des partitions. Notons qu'il a été utilisé récemment pour la recherche de partitions consensus par Green et Kreiger [GRE 99]. Plus précisément, comme nous utiliserons des variables observées quantitatives, il s'agit d'un modèle de profils latents selon la terminologie de Bartholomew et Knott [BAR 99]. On pourra consulter avec profit l'ouvrage édité par Hagenaars et McCutcheon [HAG 02].

TABLEAU 1
Les méthodes de variables latentes

	Variables latentes	
Variables observées	qualitatives	quantitatives
qualitatives	classes latentes	traits latents
quantitatives	**profils latents**	analyse factorielle

L'hypothèse de base est l'indépendance des variables observées conditionnellement aux classes latentes :

$$f(\mathbf{x}) = \sum_k \pi_k \prod_j f_k(x_j/k)$$

Les π_k sont les proportions des classes, $\mathbf{x}$ est le vecteur aléatoire des variables observées dont les composantes x_j sont indépendantes dans chaque classe, $f(x)$ est la densité de x, et $f_k(x_j/k)$ la densité de x_j dans la classe k. On sait que ce modèle souffre de problèmes sérieux d'identifiabilité, mais ici il n'est utilisé que pour engendrer des données et non pour estimer des paramètres. Il suffit alors de générer des distributions indépendantes dans chaque classe, après avoir tiré le numéro de classe de chaque observation selon une multinomiale de probabilités π_k.

3.2. L'algorithme

Pour obtenir des partitions «proches», qui ne diffèrent l'une de l'autre que de façon aléatoire, on va construire des échantillons artificiels issus d'un modèle à k classes latentes et décrits par p variables numériques, que l'on supposera par commodité normales, mais d'autres distributions sont bien sûr possibles. On partage ensuite arbitrairement les p variables en deux groupes et on effectue deux partitions en k classes des n individus selon ces deux groupes de variables à l'aide d'une méthode classique (les k-means ou nuées dynamiques.) Normalement, ces deux partitions doivent être peu différentes, on calcule alors l'indice de Rand dans toutes ses versions, l'indice dérivé de Mac Nemar, et l'indice de Jaccard.

On obtient un échantillon de valeurs de ces indices, sous l'hypothèse de «partitions proches» en itérant N fois, ce qui permet d'étudier leur distribution.

L'algorithme se déroule de la façon suivante pour une des N itérations :

- Tirage des effectifs des classes latentes selon une loi multinomiale $M(n; \pi_1, \ldots, \pi_k)$
- Pour chaque classe, tirage de p variables normales indépendantes (c'est l'hypothèse fondamentale de l'indépendance locale du modèle des classes latentes).
- Obtenir une partition $\mathcal{P}_1$ sur p_1 variables et $\mathcal{P}_2$ sur les autres $p - p_1$ variables
- Calcul des indices de comparaison de partitions.

3.3. Résultats expérimentaux

Nous avons appliqué la procédure précédente en nous limitant à 4 classes latentes équiprobables, 1000 individus et 4 variables.

Les paramètres des distributions normales ont été choisis de telle sorte que pour chaque variable j, la valeur absolue de la différence entre les moyennes de la distribution normale de deux classes différentes soit plus grande d'une fois et demie de son écart type :

$$|m_{kj} - m_{k'j}| > 1.5\sigma_j \qquad \forall j = 1, 2, 3, 4 \quad \text{et} \quad \forall k \text{ et } k' = 1, 2, 3, 4$$

m_{kj} et $m_{k'j}$ étant les moyennes respectives de la variable x_j dans les classes k et k', et σ_j l'écart type de x_j.

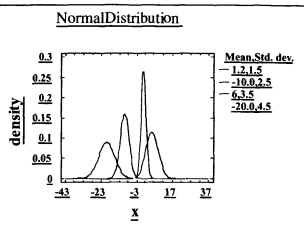
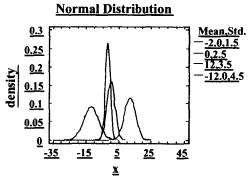
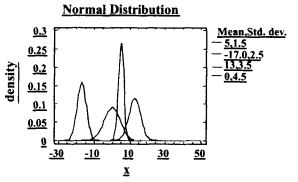
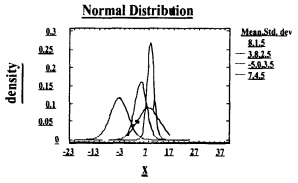
Le nombre d'itérations N vaut 1000.

Nous présentons dans la suite les résultats de nos simulations selon 2 choix de paramètres (effectuées avec le logiciel S-Plus).

3.3.1. Premier choix

Le premier choix de paramètres des 4 variables normales indépendantes pour chaque classe est présenté dans le tableau suivant :

TABLEAU 2
Les distributions par classe du premier choix

Classe 1	X_1 $N(1.2, 1.5)$ X_2 $N(-10, 2.5)$ X_3 $N(6, 3.5)$ X_4 $N(-20, 4.5)$	
Classe 2	X_1 $N(-2, 1.5)$ X_2 $N(0, 2.5)$ X_3 $N(12, 3.5)$ X_4 $N(-12, 4.5)$	
Classe 3	X_1 $N(5, 1.5)$ X_2 $N(-17, 2.5)$ X_3 $N(13, 3.5)$ X_4 $N(0, 4.5)$	
Classe 4	X_1 $N(8, 1.5)$ X_2 $N(3.8, 2.5)$ X_3 $N(-5, 3.5)$ X_4 $N(7, 4.5)$	

Pour une des 1000 itérations, la figure suivante donne la répartition spatiale dans le plan des deux premières composantes principales :

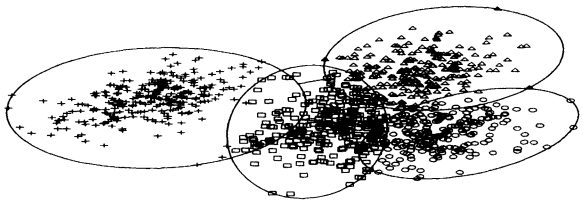


FIGURE 1
Répartition des classes du premier choix

3.3.2. Le critère symétrique de Rand

Dans la suite, nous présentons notre démarche par l'utilisation de l'indice de Rand et celui corrigé. Après avoir tiré les effectifs des classes latentes selon une loi multinomiale et tiré 4 variables normales indépendantes, on effectue deux classifications par la méthode des k-means, l'une P avec X_1 et X_2 l'autre Q avec X_3 et X_4.

On calcule les valeurs des indices de Rand R' et celui corrigé pendant ce calcul.

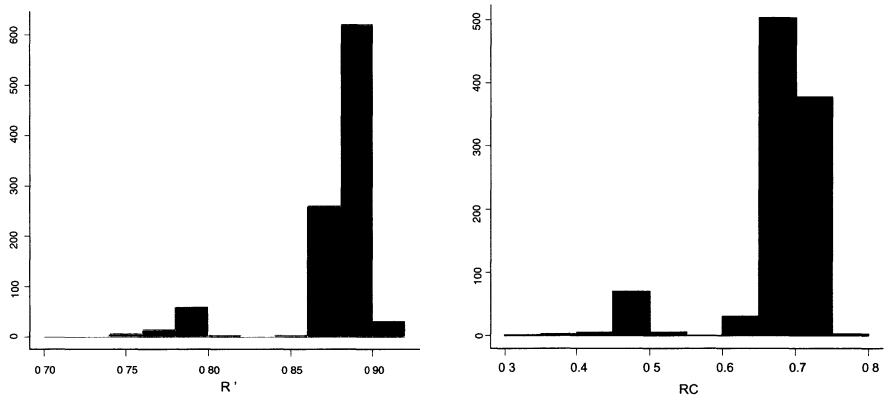


FIGURE 2
Distribution de l'indice de Rand(R') et de Rand Corrigé(RC)

Ces configurations se sont reproduites à chaque jeu de simulations, pour des paramètres différents des lois normales. Notons que dans le cas précédent, toutes les valeurs observées du coefficient de Rand sont supérieures à 0.7, alors que l'espérance de R' sous l'hypothèse d'indépendance est de 0.625, ce qui montre bien le caractère inadapté de celle-ci. Avec 1000 observations, on rejetterait l'indépendance si $R' > 0.65$ au risque de 5 % mais cela ne suffit pas pour montrer que les deux partitions sont «proches».

On ne peut cependant proposer de seuil de signification pour chacun des coefficients, car les distributions dépendent non seulement du nombre de classes, de leurs proportions mais aussi de leur séparabilité qui est liée aux paramètres des distributions normales.

On remarque que les deux indices ont une distribution de même allure (Fig. 2). Mais n'oublions pas que l'indice de Rand dans son cas corrigé peut avoir des valeurs négatives.

Nous représentons dans la figure 3 la densité de la différence $(R' - RC)$ de ces deux indices.

La différence entre l'indice de Rand R' et celui corrigé par Hubert [HUB 85] est très proche d'une distribution normale de moyenne de 0.19.

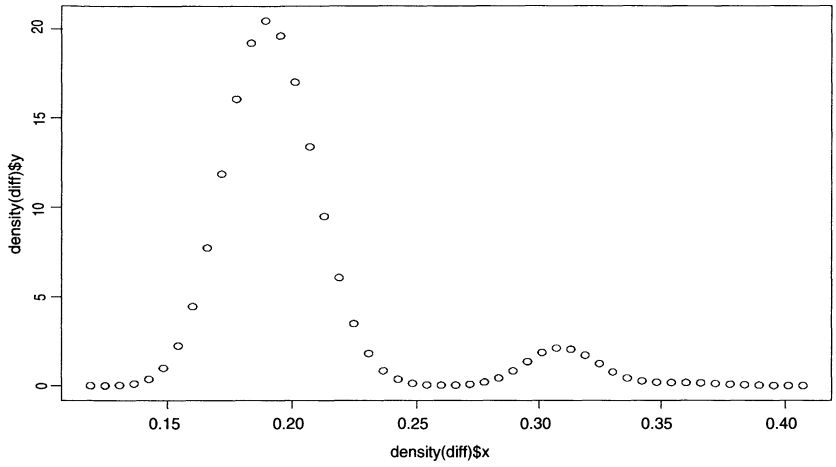


FIGURE 3
Densité de la différence entre l'indice de RandR' et celui corrigé

3.3.3. Les indices dérivés de Mc Nemar et de Jaccard

Comme l'indice de Rand donne la même importance aux couples d'individus qui sont dans la même classes de deux partitions, qu'à ceux qui ne sont pas dans la même classe pour les deux partitions (accord négatif), on utilise la même démarche en calculant les indices de Mc Nemar et celui de Jaccard pour 1000 simulations.

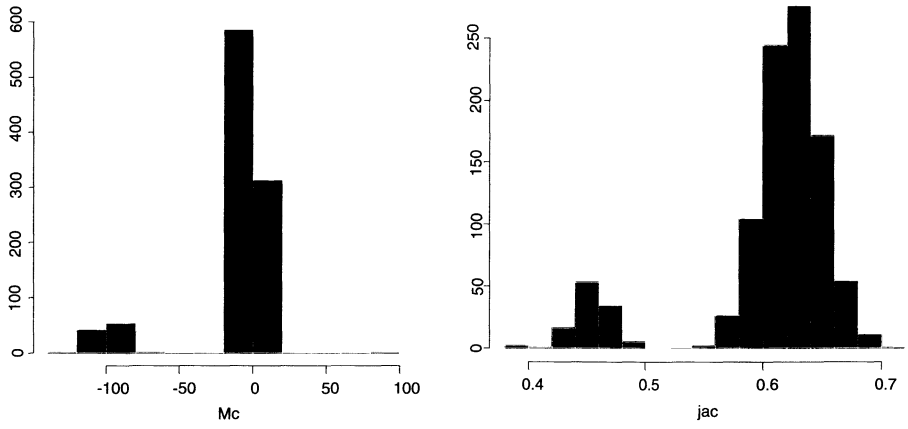


FIGURE 4
Distribution de l'indice de Mc Nemar et de Jaccard

L'indice de Mc Nemar est à majorité distribué autour de zéro montrant ainsi que pour un risque de 5 % l'hypothèse nulle est vérifiée. La distribution de l'indice de Jaccard présente des valeurs supérieures à 0.4 dont la valeur la plus fréquente est de 0.63. On remarque encore ici une bimodalité surprenante.

3.3.4. Le critère asymétrique de Rand

La même procédure est utilisée pour trouver les variables normales indépendantes, en effectuant deux autres classifications : la première partition $\mathcal{P}_1$ de X_1, X_2 et X_3 formée de 6 classes, et la deuxième partition $\mathcal{P}_2$ de X_4 formée de 3 classes par la méthode des k-means.

Dans ce cas, on cherche à évaluer dans quelle mesure les classes de $\mathcal{P}_1$ sont incluses dans celles de $\mathcal{P}_2$. On calcule alors l'indice de Rand asymétrique RA' et celui corrigé aussi 1000 fois.

On trouve les résultats suivants :

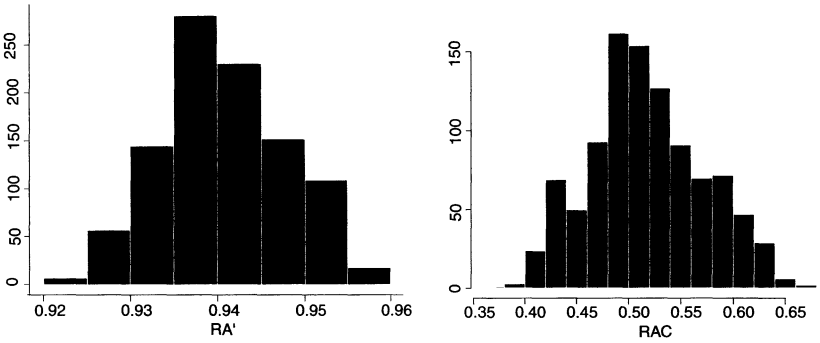


FIGURE 5
Distribution de l'indice de Rand asymétrique et de Rand corrigé

On remarque que les valeurs de l'indice de Rand asymétrique RA' sont supérieures à 0.92. Par contre celui de Rand asymétrique corrigé prend ses valeurs à partir de 0.36.

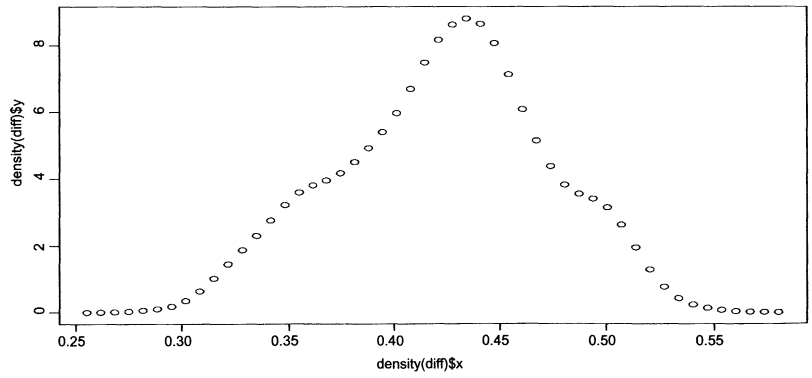


FIGURE 6
Répartition de la densité de la différence de l'indice de Rand asymétrique RA' et de Rand corrigé RA_C

La représentation graphique de la densité de la différence entre l'indice de Rand asymétrique et celui corrigé $(RA' - RA_c)$ est donnée dans la figure 6.

La différence de RA' et RA_C est approximativement une distribution normale de moyenne 0.435.

Afin de comparer l'indice de Rand brut R' et celui de Rand asymétrique RA', on représente la distribution de Rand brut R' pour cette même partition (Figure 7) :

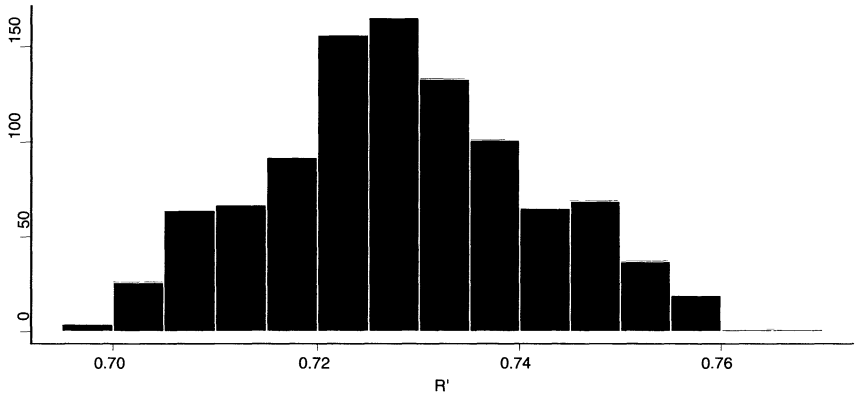


FIGURE 7
Distribution de l'indice de Rand R' des partitions asymétriques

Contrairement à ce qu'on a trouvé dans les résultats des partitions symétriques, on a une distribution modale dans tous les cas de l'indice de Rand. Cela revient à conclure que ces distributions dépendent du nombre de classes dans chaque partition.

3.4. Deuxième choix de paramètre

Le graphique suivant dans le plan des deux premières composantes principales montre la répartition spatiale d'une des 1000 itérations.

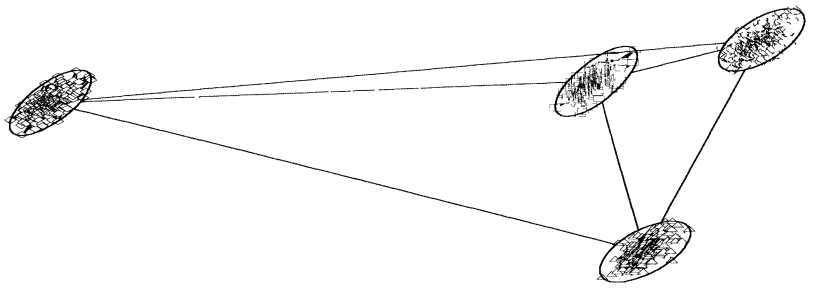
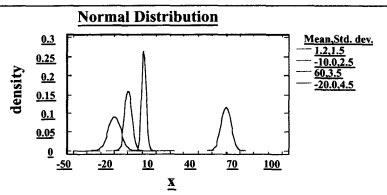
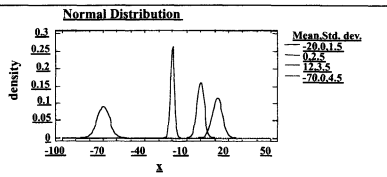
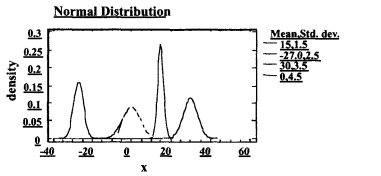
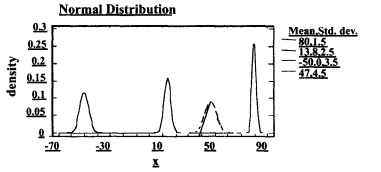


FIGURE 8
Répartition des classes du deuxième choix

TABLEAU 3
Les distributions par classe du deuxième choix

Classe 1	X_1 $N(1.2, 1.5)$ X_2 $N(-10, 2.5)$ X_3 $N(60, 3.5)$ X_4 $N(-20, 4.5)$	
Classe 2	X_1 $N(-20, 1.5)$ X_2 $N(0, 2.5)$ X_3 $N(12, 3.5)$ X_4 $N(-70, 4.5)$	
Classe 3	X_1 $N(15, 1.5)$ X_2 $N(-27, 2.5)$ X_3 $N(30, 3.5)$ X_4 $N(0, 4.5)$	
Classe 4	X_1 $N(80, 1.5)$ X_2 $N(13.8, 2.5)$ X_3 $N(-50, 3.5)$ X_4 $N(47, 4.5)$	

Afin de tester l'algorithme dans différents choix de paramètres de distributions normales, on a choisi un cas où les classes sont nettement séparées. On trouve une distribution de Rand R' non bimodale (Figure 9) et de valeur égale à 1. Il est clair que la distribution de l'indice de Rand dépend de la séparation des classes, de nombre d'individus et du nombre de classes.

3.5. Autres choix de paramètres

On présente maintenant d'autres simulations de l'indice de Rand R' avec 4 nouveaux choix de paramètres des distributions normales pour les 4 variables à 4 classes latentes équiprobables. Les deux partitions P_1 de X_1, X_2, et P_2 de X_3, X_4 sont formées toujours par la méthodes de k-means. Le nombre d'itérations N vaut 500.

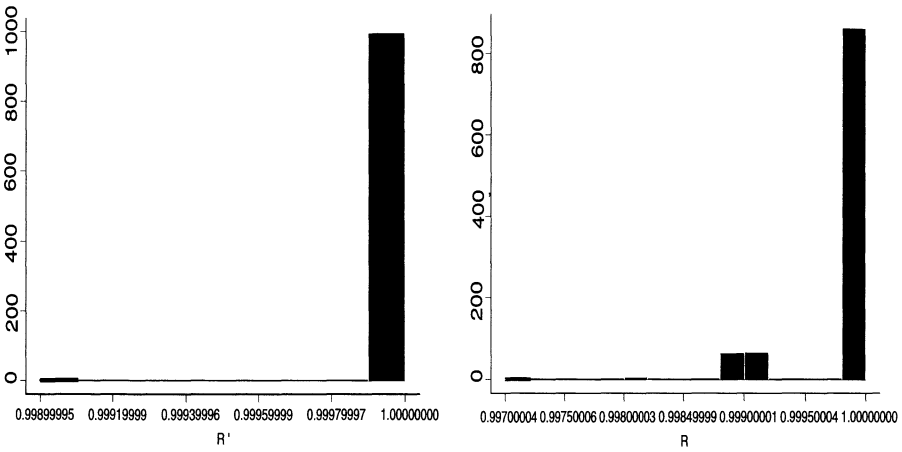
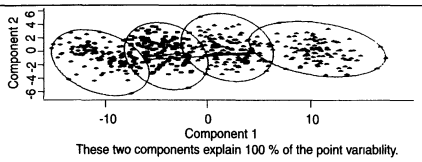
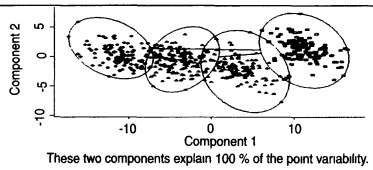
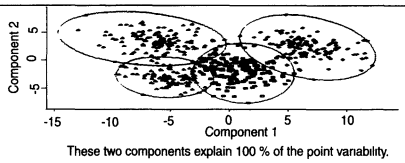
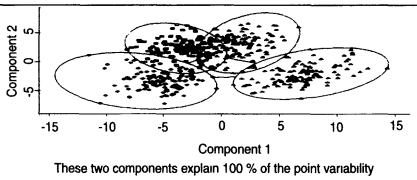


FIGURE 9
Distribution de l'indice de Rand brut R' et R selon le deuxième choix de paramètres

TABLEAU 4
Le choix de paramètres par classes pour 4 nouveaux cas

Cas	Classe 1	Classe 2	Classe 3	Classe 4
Cas 1	X_1 $N(1, 1.5)$ X_2 $N(4.2, 2.5)$ X_3 $N(7, 3.5)$ X_4 $N(10, 4.5)$	X_1 $N(-2, 1.5)$ X_2 $N(-4, 2.5)$ X_3 $N(-6, 3.5)$ X_4 $N(-10, 4.5)$	X_1 $N(-5, 1.5)$ X_2 $N(-10, 2.5)$ X_3 $N(-13, 3.5)$ X_4 $N(-20, 4.5)$	X_1 $N(-8, 1.5)$ X_2 $N(-15, 2.5)$ X_3 $N(-20, 3.5)$ X_4 $N(-30, 4.5)$
Cas 2	X_1 $N(1, 1.5)$ X_2 $N(4, 2.5)$ X_3 $N(7, 3.5)$ X_4 $N(10, 4.5)$	X_1 $N(-2, 1.5)$ X_2 $N(-4, 2.5)$ X_3 $N(-6, 3.5)$ X_4 $N(-10, 4.5)$	X_1 $N(-5, 1.5)$ X_2 $N(-10, 2.5)$ X_3 $N(-13, 3.5)$ X_4 $N(-20, 4.5)$	X_1 $N(8, 1.5)$ X_2 $N(8.2, 2.5)$ X_3 $N(20, 3.5)$ X_4 $N(30, 4.5)$
Cas 3	X_1 $N(1, 1.5)$ X_2 $N(4.2, 2.5)$ X_3 $N(7, 3.5)$ X_4 $N(10, 4.5)$	X_1 $N(4, 1.5)$ X_2 $N(8, 2.5)$ X_3 $N(1.5, 3.5)$ X_4 $N(3, 4.5)$	X_1 $N(7, 1.5)$ X_2 $N(0, 2.5)$ X_3 $N(13, 3.5)$ X_4 $N(17, 4.5)$	X_1 $N(10, 1.5)$ X_2 $N(12, 2.5)$ X_3 $N(18.5, 3.5)$ X_4 $N(24, 4.5)$
Cas 4	X_1 $N(4, 1.5)$ X_2 $N(20, 2.5)$ X_3 $N(17.8, 3.5)$ X_4 $N(37, 4.5)$	X_1 $N(11, 1.5)$ X_2 $N(16, 2.5)$ X_3 $N(12.3, 3.5)$ X_4 $N(17, 4.5)$	X_1 $N(13.5, 1.5)$ X_2 $N(12, 2.5)$ X_3 $N(1.5, 3.5)$ X_4 $N(3, 4.5)$	X_1 $N(8, 1.5)$ X_2 $N(8.2, 2.5)$ X_3 $N(20, 3.5)$ X_4 $N(30, 4.5)$

TABLEAU 5
La moyenne de l'indice de Rand R' selon les distributions par classe de 4 nouveaux choix

Cas	Moyenne de R'	Répartitions spatiales
Cas 1	0.9633996	
Cas 2	0.9110861	
Cas 3	0.775902	
Cas 4	0.8026252	

En effectuant 500 itérations, l'indice de Rand R' prend des valeurs de moyennes 0.9 même si la séparation des classes n'est pas très grande (cas 1 et 2) donc cet indice dépend de plus du nombre de classe et du nombre d'individus.

3.6. Variations du nombre de classes

Pour 1000 itérations, et pour un type de choix de paramètres, on cherche la moyenne de l'indice de Rand R', de Mc Nemar et de Jaccard, en faisant varier le nombre de classes k de 3 à 5. On trouve les résultats suivants :

TABLEAU 6
Moyennes des indices selon le nombre de classes latentes k en 1000 itérations

N	n	k	Moyenne de Rand R'	Moyenne de Jaccard	Moyenne de Mc Nemar
1000	1000	3	0.8602054	0.6567402	−5.099624
1000	1000	4	0.8562098	0.5556003	−5.903291
1000	1000	5	0.8916331	0.585066	0.0709304

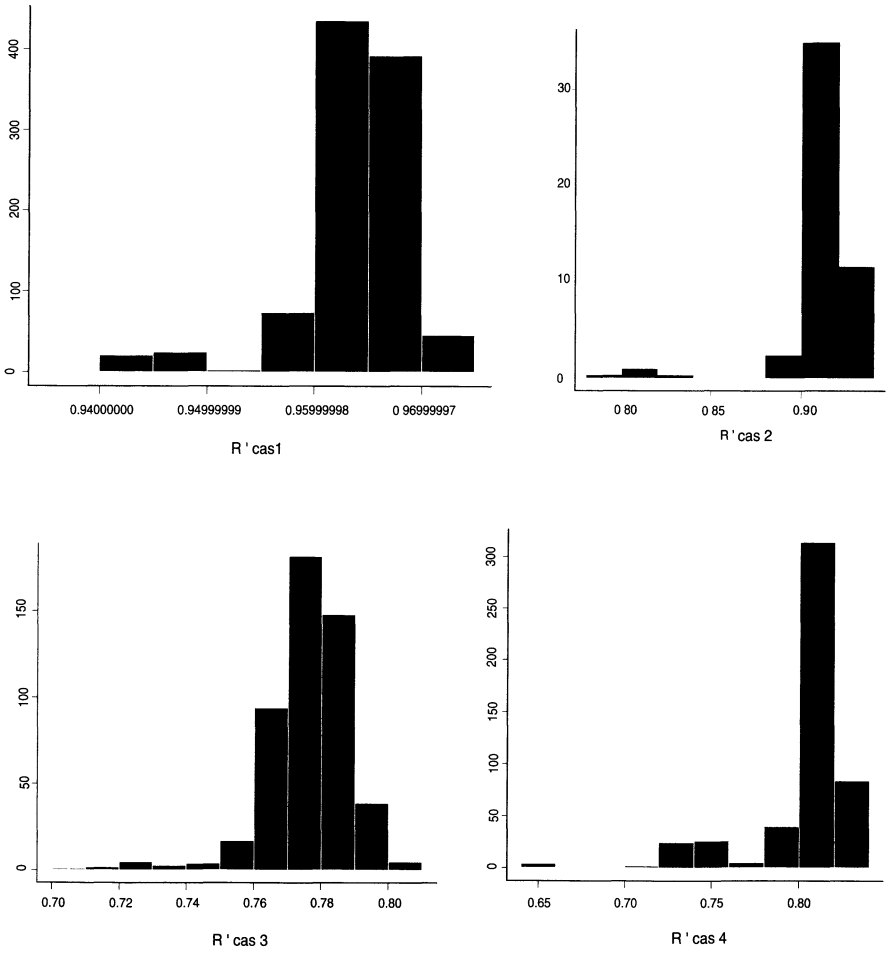


FIGURE 10
Distribution de l'indice de Rand brut R' selon les nouveaux choix de paramètres

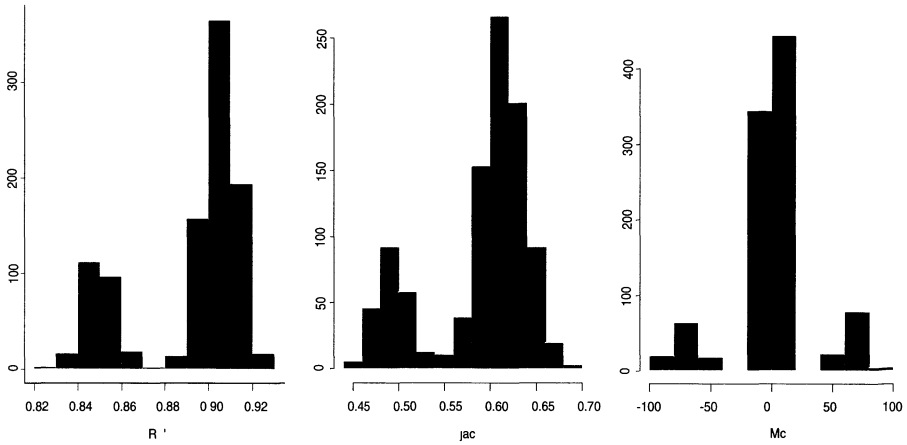


FIGURE 11
Distribution de l'indice de Rand brut R', de l'indice de Jaccard et de Mc Nemar dans le cas de 5 classes

Les moyennes de l'indice de Rand R' et de l'indice dérivé de Mc Nemar croissent lorsqu'on augmente le nombre de classes latentes équiprobables. Par contre l'indice de Jaccard varie en sens inverse du nombre de classes. Cela peut être normal, car en augmentant le nombre de classes, les paires d'individus (i, i') qui sont dans une même classe dans la première partition ont peu de chance de rester ensemble dans la deuxième partition.

Pour un cas choisi du nombre de classe ($k = 5$), on trouve la bimodalité de la distribution de Rand à partir de 0.86. Les valeurs observées de l'indice de Jaccard sont supérieures à 0.45. Pourtant les valeurs les plus fréquentes de l'indice de Mc Nemar sont à zéro.

3.7. Comparaison entre la moyenne théorique et expérimentale de l'indice de Rand R'

L'espérance théorique de l'indice de Rand selon Idrissi sous l'hypothèse d'indépendance est donnée par la valeur suivante :

$$E(R') = 1 - \frac{2}{k} + \frac{2}{k^2}$$

En changeant le nombre de classes k d'une partition, on trouve pour 1000 itérations les moyennes de l'indice de Rand par simulation en utilisant l'algorithme proposé :

TABLEAU 7
Moyenne théorique et moyenne expérimentale de l'indice de Rand R' selon le nombre de classes k en 1000 itérations

n	k	$E(R')$	M_{exp} de R'	$\Delta = \lvert M_{\text{exp}} - E(R')\rvert$
1000	4	0.625	0.8562098	0.2312098
1000	5	0.68	0.8916331	0.2116331
1000	6	0.72	0.8833446	0.1633446
1000	7	0.775	0.8745753	0.0995753
1000	8	0.78125	0.8802063	0.0989563

La moyenne de Rand théorique croît avec le nombre de classes, mais ce n'est pas toujours le cas pour la moyenne expérimentale trouvée par simulation.

La différence entre les valeurs théoriques et expérimentales croit de -23 % à -9 % lorsque le nombre de classes des partitions augmente, mais elle décroît en valeur absolue.

3.8. Sur la bimodalité

On a pu remarquer le caractère bimodal de la distribution du coefficient de Rand brut R' avec un mode secondaire correspondant à environ 10 % des cas lorsque la séparation des classes n'est pas grande. Ce phénomène peut en fait s'expliquer par le caractère non-optimal de la méthode de classification utilisée : on sait que les k-means fournissent une solution dépendant du choix initial des centres. Or la procédure utilisée dans S+ fait une classification hiérarchique ascendante avant de lancer les k-means et part toujours de la même initialisation obtenue à partir d'une coupure en k classes de l'arbre hiérarchique. En modifiant le choix initial des centres par la procédure FASTCLUS de SAS nous avons pu obtenir des partitions finales différentes (souvent meilleures au sens de l'inertie) et une augmentation de l'indice de Rand. En appliquant une deuxième fois la méthode des k-means aux 10 % de valeurs correspondant au mode secondaire, on a pu obtenir une augmentation sensible de R'.

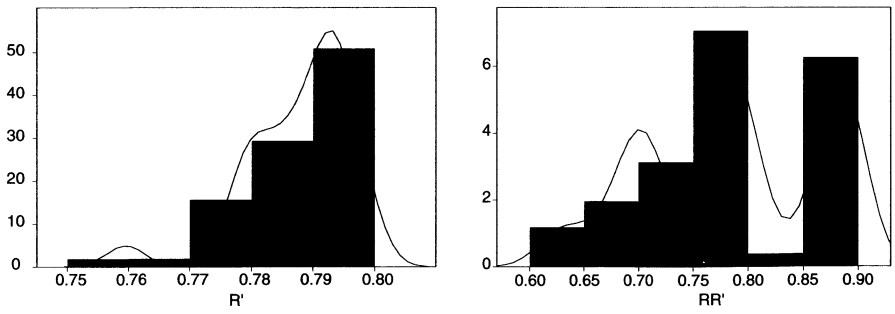


FIGURE 12
Distribution de l'indice de Rand R' en appliquant une et deux fois les k-means

4. Discussion

L'utilisation d'un modèle de classes latentes a permis d'aborder le problème de la proximité de deux partitions. Celle-ci peut-être mesurée par l'indice de Rand brut ou corrigé, ainsi que par sa version asymétrique et par les deux indices dérivés de Mac Nemar et celui de Jaccard.

On a fait deux choix de paramètres des variables normales pour comparer les résultats trouvés.

L'indice de Rand dans sa version utilisée, donne la même importance aux couples d'individus qui sont dans la même classe dans les deux partitions, qu'à ceux qui ne sont pas dans la même classe pour les deux partitions (accord «négatif»), ce qui est contestable. D'où l'utilisation des autres indices de comparaison. L'indice de Mc Nemar s'intéresse aux couples d'individus qui ont changé de classes (désaccord). L'indice de Jaccard mesure l'accord positif des partitions.

La distribution de ces indices, qui est très différente de celle obtenue sous l'hypothèse d'indépendance, n'a été étudiée que par simulation et dans des cas particuliers. Il est clair que ces distributions dépendent du nombre de classes des partitions, des nombres d'individus et de la plus ou moins grande séparation des classes, etc. ce qui empêche de donner des bornes universelles. Des études par bootstrap ou permutations aléatoires permettraient de trouver des bornes *ad hoc*.

Références

[BAR 99] BARTHOLOMEW D. J., KNOTT M., *Latent Variable Models and Factor Analysis*, Arnold, London, 1999.

[CHAV 01] CHAVENT M. *et al.*, «Critère de Rand asymétrique», *Congrès de la SFC*, Pointe à Pitre, décembre 2001.

[DAY 83] DAY W.H.E, «The Role of Complexity in Comparing Classifications», *Mathematical Biosciences*, 66, 97- 114, 1983.

[GRE 99] GREEN P., KREIGER A., «A Generalized Rand-Index Method for Consensus Clustering of Separate Partitions of the Same Data Base», *Journal of Classification*, 16, 63-89,1999.

[HAG 02] HAGENAARS J.A., MCCUTCHEON A.I. editors, *Applied Latent Class Analysis*, Cambridge University Press, 2002

[HUB 85] HUBERT L., ARABIE P., «Comparing partitions», *Journal of Classification*, 2, 193-198, 1985.

[IDR 00] IDRISSI A., *Contribution à l'unification de critères d'association pour variables qualitatives*, Thèse de doctorat de l'Université de Paris 6, 2000.

[KEN 61] KENDALL M-G, STUART A., *The Advanced Theory of Statistics*, Vol 2, Griffin, Londres, 1961.

[MAR 84] MARCOTORCHINO J.F., *Utilisation des Comparaisons par Paires en Statistique des Contingences (Partie II)*, Étude du Centre Scientifique IBM France, No F069, 1984.

[MAR 91] MARCOTORCHINO J.F., EL AYOUBI N., «Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association», *Revue de Statistique Appliquée*, XXXIX , 2, 25-46, 1991.

[MILL 86] MILLIGAN G.W., COOPER M.C., «A study of the comparability of external criteria for hierarchical cluster Analysis», *Multivariate Behavior Research*, vol. 21, 441-458, 1986.

[SAP 97] SAPORTA G., «Problèmes Posés par la Comparaison de Classifications Dans des Enquêtes Différentes», *53ème session de l'Institut International de Statistique*, Istanbul, août 1997.

[SAP 02] SAPORTA G., YOUNESS G., «Comparing two partitions : some proposals and Experiments», *Proceedings in Computational Statistics*, W. Härdle (ed.), Physica-Verlag, Berlin, 2002.