



Mathématiques et sciences humaines

Mathematics and social sciences

154 | Été 2001

Analyse statistique implicative

Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé

Maximizing association by grouping rows or columns of a crosstable

Gilbert Ritschard, Djamel Zighed et Nicolas Nicoloyannis



Édition électronique

URL : <http://journals.openedition.org/msh/2841>

DOI : 10.4000/msh.2841

ISSN : 1950-6821

Éditeur

Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée

Date de publication : 1 mars 2001

ISSN : 0987-6936

Référence électronique

Gilbert Ritschard, Djamel Zighed et Nicolas Nicoloyannis, « Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé », *Mathématiques et sciences humaines* [En ligne], 154 | Été 2001, mis en ligne le 10 février 2006, consulté le 23 juillet 2020. URL : <http://journals.openedition.org/msh/2841> ; DOI : <https://doi.org/10.4000/msh.2841>

MAXIMISATION DE L'ASSOCIATION PAR REGROUPEMENT DE LIGNES OU DE COLONNES D'UN TABLEAU CROISÉ

Gilbert RITSCHARD*, Djamel A. ZIGHED et Nicolas NICOLOYANNIS†

RÉSUMÉ – *L'intensité de l'association entre la variable ligne et la variable colonne d'un tableau croisé varie avec le regroupement de catégories. Dans plusieurs contextes, comme la discrétisation simultanée de deux variables, il importe de déterminer le niveau de regroupement qui maximise l'association. Les principales mesures d'association suite à une agrégation de lignes ou de colonnes sont étudiées et une heuristique permet de déterminer le regroupement qui (quasi-)maximise le degré d'association. Des simulations comparant les quasi-optima aux vrais optima servent à évaluer la fiabilité de l'algorithme proposé.*

MOTS CLÉS – Table de contingence, Agrégation, Association, Discrétisation

SUMMARY – *Maximizing association by grouping rows or columns of a crosstable. The strength of association between the row and column variables in a crosstable varies with the level of aggregation of each variable. In many settings such as the simultaneous discretization of two variables, it is useful to determine the aggregation level that maximizes the association. The main association measures with respect to the aggregation of rows and columns are studied and permits a heuristic algorithm to (quasi-)maximize the association through aggregation. Simulations carried out to investigate the reliability of the algorithm are presented.*

KEYWORDS – Crosstable, Aggregation, Association, Discretization

1 INTRODUCTION

L'étude de l'association entre variables catégorielles repose en général sur une analyse du tableau croisé des variables concernées. On commence par tester l'indépendance en examinant les statistiques du χ^2 de Pearson ou du rapport de vraisemblance associées au tableau, puis, pour saisir l'intensité de l'association on se réfère à des mesures

*Département d'économétrie, Université de Genève, bd du Pont-d'Arve 40, CH-1211 Genève 4, e-mail : gilbert.ritschard@themes.unige.ch

†Laboratoire ERIC de l'Université Lumière Lyon 2, e-mail : (zighed,nicolas.nicoloyannis)@univ-lyon2.fr. L'essentiel de ce travail a été réalisé pendant les séjours de G. Ritschard en qualité de professeur invité au laboratoire ERIC en 2000 et 2001.

d'association telles que le v de Cramer, le τ ou le γ de Goodman et Kruskal, le coefficient d'incertitude de Theil, le τ_b de Kruskal, le d de Somers, qui sont elles-mêmes calculées sur la base de la table de contingence.

La question abordée ici est celle de l'effet de l'agrégation de lignes ou colonnes d'un tableau croisé sur les mesures d'association. Il est bien connu par exemple que le regroupement de colonnes ou de lignes qui ont la même distribution conditionnelle n'affecte pas les statistiques du χ^2 de Pearson et du rapport de vraisemblance (voir par exemple [2] ou [16], p. 450) mais renforce en général le degré d'association comme l'illustre en particulier les simulations présentées dans [17]. Notre propos est d'étudier ce lien entre fusion de lignes ou de colonnes et degré d'association dans le but de déterminer le regroupement optimal, à savoir celui qui donne lieu à la plus forte association entre les variables ligne et colonne.

La maximisation du degré d'association trouve sa motivation dans plusieurs domaines. Par exemple, l'analyse de données collectées par questionnaires nécessite en général, pour des raisons d'effectifs notamment, un regroupement a posteriori des items en catégories. Lorsqu'il s'agit, comme c'est souvent le cas en sciences sociales, d'étudier l'association entre variables, il est utile de comprendre les effets du regroupement sur l'association et de choisir le cas échéant celui qui rend le mieux compte du lien. Une seconde motivation a trait à la discrétisation qui est en particulier un problème majeur en apprentissage. Les solutions optimales par rapport à la discrimination recherchée que propose la littérature et répertoriées par exemple dans [19] procèdent individuellement pour chaque variable. Dans une optique prévisionnelle, des solutions bidimensionnelles, où l'on discrétise conjointement deux variables, devraient s'avérer plus efficaces. Breiman et al. [7] ont étudié le cas de la dichotomisation conjointe de deux variables. Notre propos est de généraliser ce cas à un nombre quelconque de catégories.

Hormis les cas triviaux avec un nombre initial réduit de l'ordre de trois ou quatre catégories, rechercher la solution optimale par recombinaison systématique ne peut pas être envisagé à cause du nombre exponentiel de cas à considérer. Il s'agit alors de formuler une procédure qui puisse être exploitée algorithmiquement pour trouver les regroupements conjoints des catégories ligne et colonne qui (quasi-)maximisent un critère donné d'association. Notons qu'il conviendra ici de distinguer le cas des variables nominales de celui des variables ordinales pour lesquelles seuls des regroupements de catégories adjacentes ont un sens.

L'optimisation considérée n'est pas sans rappeler la question des partitions des lignes et des colonnes qui maximisent le χ^2 de Pearson, abordée par Benzécri [4], et traitée en particulier par Celeux et al. [8] au moyen d'une algorithmique de nuées dynamiques appliquée alternativement sur les lignes et les colonnes. Ces auteurs se placent cependant dans un contexte où, contrairement à celui retenu ici, seules des partitions avec un nombre de classes fixé a priori sont pertinentes. Plus généralement, la recherche simultanée de partitions optimales des lignes et des colonnes d'une matrice a été considérée par exemple par W.D. Fisher [9] [10] dans le but de simplifier les modèles économiques de prévision et de décision à équations multiples. Le problème a également été étudié pour des matrices de données où il s'agit de partitionner simultanément les cas et les variables, voir par exemple [1], [6] et [13]. Le cas particulier de

tableaux binaires a notamment été traité par Govaert [12] [13]. En ce qui concerne plus précisément les tables de contingence, l'agrégation ou le partitionnement simultané des lignes et des colonnes a été examiné de plusieurs points de vue. En plus de celui de Benzécri déjà mentionné on peut citer Gilula et Krieger [11] qui étudient le comportement du χ^2 de Pearson suite à la réduction de tables par agrégation, ainsi que Hirotsu [15] et Greenacre [14] qui mettent en évidence des sous-tables homogènes. Si ces travaux s'apparentent à notre problématique par l'aspect de l'agrégation simultanée des lignes et des colonnes, ils s'en distinguent sur les deux points suivants : le critère d'optimisation qui est par exemple l'homogénéité de sous-tables pour Hirotsu et Greenacre tandis que nous nous intéressons explicitement au degré d'association ; les dimensions finales du tableau qui sont fixées *a priori* par exemple chez Benzécri, Celeux et al. et Govaert, alors que notre objectif est de trouver les dimensions optimales.

Dans la section 2, nous illustrons avec un exemple numérique la variété des effets d'une agrégation de catégories sur les statistiques du χ^2 et les mesures d'association. Le cadre formel et les notations sont précisées en 3. La section 4 précise le degré de complexité de la recherche de la solution optimale et introduit le principe d'une heuristique pas à pas. La section 6 propose une étude analytique de l'effet du regroupement de deux catégories ligne ou colonne sur un choix de mesures d'association nominales et ordinales. Il s'agit d'une part d'obtenir dans la mesure du possible des expressions simples de la variation des critères d'association, et d'autre part de mieux comprendre la sensibilité des mesures d'association, en particulier dans le cas de l'équivalence distributionnelle. La section 7 rapporte les résultats d'une étude par simulations des écarts entre le quasi-optimum fourni par l'heuristique proposée et l'optimum global. Nous concluons à la section 8 avec des perspectives de développement.

2 EXEMPLE ET RÉSULTATS INTUITIFS

Considérons le tableau croisé suivant entre une variable ligne x et une variable colonne y :

$$M = \begin{array}{c|cccc} x \backslash y & A & B & C & D \\ \hline a & 10 & 10 & 1 & 1 \\ b & 10 & 10 & 1 & 1 \\ c & 1 & 1 & 10 & 10 \\ d & 1 & 1 & 10 & 10 \end{array}$$

Intuitivement, un regroupement des deux premières $\{A, B\}$ ou des deux dernières colonnes $\{C, D\}$, qui sont identiques, devraient renforcer le degré d'association. Il en est de même pour un regroupement des lignes $\{a, b\}$ ou $\{c, d\}$. Par contre, un regroupement des catégories B et C par exemple, diminue le contraste entre les distributions colonnes et devrait donc se traduire par une réduction de l'association.

Pour illustrer ces aspects on considère alors d'une part les tableaux :

$$M_y^+ = \begin{array}{c|ccc} x \backslash y & A & B & \{C, D\} \\ \hline a & 10 & 10 & 2 \\ b & 10 & 10 & 2 \\ c & 1 & 1 & 20 \\ d & 1 & 1 & 20 \end{array} \quad M_{xy}^+ = \begin{array}{c|ccc} x \backslash y & A & B & \{C, D\} \\ \hline a & 10 & 10 & 2 \\ b & 10 & 10 & 2 \\ \{c, d\} & 2 & 2 & 40 \end{array}$$

et d'autre part les tableaux :

$$M_y^- = \begin{array}{c|ccc} x \backslash y & A & \{B, C\} & D \\ \hline a & 10 & 11 & 1 \\ b & 10 & 11 & 1 \\ c & 1 & 11 & 10 \\ d & 1 & 11 & 10 \end{array} \quad M_{xy}^- = \begin{array}{c|ccc} x \backslash y & A & \{B, C\} & D \\ \hline a & 10 & 11 & 1 \\ \{b, c\} & 11 & 22 & 11 \\ d & 1 & 11 & 10 \end{array}$$

Tableau 1 : Mesures d'association selon le regroupement

	M	M_y^+	M_{xy}^+	M_y^-	M_{xy}^-
Lignes	4	4	3	4	3
Colonnes	4	3	3	3	3
Degrés liberté	9	6	4	6	4
Khi-2 Pearson	58.91	58.91	58.91	29.45	14.73
Rapport vraisemblance	68.38	68.38	68.38	34.19	17.09
t Tschuprow	0.47	0.52	0.58	0.37	0.29
v Cramer	0.47	0.58	0.58	0.41	0.29
$\tau_{y \leftarrow x}$ Goodman-Kruskal	0.22	0.40	0.40	0.13	0.07
$\tau_{x \leftarrow y}$ Goodman-Kruskal	0.22	0.22	0.40	0.11	0.07
$u_{y \leftarrow x}$ Coefficient incertitude	0.28	0.37	0.37	0.19	0.09
$u_{x \leftarrow y}$ Coefficient incertitude	0.28	0.28	0.37	0.14	0.09
γ Goodman-Kruskal	0.68	0.77	0.80	0.63	0.57
τ_b Kendall	0.55	0.60	0.65	0.45	0.37
$d_{y \leftarrow x}$ Somers	0.55	0.55	0.65	0.41	0.37
$d_{x \leftarrow y}$ Somers	0.55	0.65	0.65	0.49	0.37

Le tableau 1 résume les valeurs d'un choix de mesures d'association pour le tableau M et les quatre regroupements considérés. On observe que si, conformément au principe de l'équivalence distributionnelle, les mesures du χ^2 restent les mêmes pour les tableaux M , M_y^+ et M_{xy}^+ , les mesures d'association augmentent par contre comme pressenti avec l'agrégation de lignes et de colonnes semblables. Les valeurs pour les regroupements M_y^- et M_{xy}^- font apparaître que l'agrégation de colonnes ou de lignes distribuées de façon très différentes conduit à une réduction aussi bien des statistiques du χ^2 que des mesures d'association.

On a ainsi comme premières indications que l'association :

- s'amplifie par le regroupement de catégories d'une variable dont les effectifs se distribuent de façon similaire selon l'autre variable ;

- *s'atténue* par le regroupement de catégories d'une variable dont les effectifs se distribuent de façon très différente selon l'autre variable.

3 CADRE FORMEL ET NOTATIONS

Soit deux variables x et y prenant respectivement ℓ et c états différents. Le croisement des variables donne lieu à une table de contingence $T_{\ell \times c}$ ℓ lignes et c colonnes. On note n_{ij} l'effectif de la cellule de la table se trouvant à l'intersection de la i -ème ligne et de la j -ème colonne. Les totaux des lignes et des colonnes sont représentés en remplaçant l'indice de sommation par un point : $n_{i.} = \sum_j n_{ij}$, $n_{.j} = \sum_i n_{ij}$. Enfin, on note n l'effectif total de la table : $n = \sum_i n_{i.} = \sum_j n_{.j}$. Ceci concerne évidemment les cas observés. Le cas échéant, on se référera aux probabilités p_{ij} , $p_{i.}$ et $p_{.j}$ qu'un individu choisi au hasard dans la population de référence soit respectivement dans la case (i, j) , la ligne i ou la colonne j .

On considère les critères d'association θ_{xy} entre x et y en tant que fonction des éléments de la table de contingence $\theta_{xy} = \theta(T_{\ell \times c})$.

Soit P_x une partition des états de la variable ligne x et P_y une partition des valeurs de y . Chaque couple (P_x, P_y) de partitions donne lieu à une table de contingence T différente. Le problème général envisagé est alors la recherche du couple de partitions qui maximise la mesure d'association :

$$\max_{P_x, P_y} \theta(T(P_x, P_y)) \quad (1)$$

Pour des variables ordinales, et donc en particulier pour les variables mesurables de type intervalle ou ratio, seules les partitions obtenues par regroupement de catégories adjacentes sont pertinentes. Dans ce cas, on considère le problème restreint :

$$\begin{cases} \max_{P_x, P_y} \theta(T(P_x, P_y)) \\ \text{s.c. } P_x \in \mathcal{A}_x \text{ et } P_y \in \mathcal{A}_y \end{cases} \quad (2)$$

où \mathcal{A}_x et \mathcal{A}_y désignent respectivement l'ensemble des partitions par regroupements adjacents des catégories de x et de y . En désignant par \mathcal{P}_x et \mathcal{P}_y les ensembles non restreints de partitions, on a, pour $c, \ell > 2$, $\mathcal{A}_x \subset \mathcal{P}_x$ et $\mathcal{A}_y \subset \mathcal{P}_y$. L'association entre variables ordinales pouvant être négative, le critère $\theta(T(P_x, P_y))$ à maximiser est dans ce cas la valeur absolue d'une mesure d'association.

4 STRATÉGIE GÉNÉRALE

Précisons le degré de complexité de la recherche de la solution optimale par exploration et décrivons le principe de l'heuristique qui sera exploitée. La complexité d'une exploration complète empêche sa généralisation à un nombre quelconque de lignes et de colonnes et justifie en effet le recours à une heuristique.

4.1 COMPLEXITÉ DE LA SOLUTION OPTIMALE

La détermination de la solution optimale nécessite l'exploration, de tous les regroupements possibles de lignes ou de colonnes, c'est-à-dire l'ensemble des couples (P_x, P_y) . Le nombre de situations à examiner est donné par le produit du nombre de regroupements possibles des lignes par celui des colonnes, soit $\#\mathcal{P}_x \#\mathcal{P}_y$.

Pour le cas de variables nominales, le nombre de regroupements correspond au nombre $B(c) = \#\mathcal{P}$ de partitions de l'ensemble des c catégories de la variable. Il peut être obtenu récursivement par la formule de Bell [3]

$$B(c) = \sum_{0 \leq k \leq c-1} \binom{c-1}{k} B(k)$$

avec $B(0) = 1$. Pour $c = \ell$ le nombre $B(c)B(\ell)$ de configurations possibles à explorer est ainsi par exemple respectivement 25, 225, 2704 et 41209 pour $c = 3, 4, 5, 6$ et passe à plus de 13 milliards pour $c = \ell = 10$.

Pour des variables ordinales, et donc en particulier dans les problèmes de discrétisation, seuls les regroupements de catégories adjacentes sont pertinentes. Le nombre de cas à explorer s'en trouve dès lors réduit. Le nombre $G(c) = \#\mathcal{A}$ de groupements différents de c catégories est :

$$G(c) = \sum_{k=0}^{c-1} \binom{c-1}{k} = 2^{(c-1)}$$

ce qui donne respectivement $G(c)G(\ell) = 16, 64, 256, 1024$ configurations à explorer pour un tableau carré avec $c = \ell = 3, 4, 5, 6$ et plus d'un million pour $c = \ell = 10$.

4.2 HEURISTIQUE PAS À PAS

Compte tenu des limites évidentes de l'exploration exhaustive de toutes les configurations, nous procédons par regroupements successifs de deux catégories. Cette façon de faire ne conduit évidemment pas nécessairement à la solution optimale, mais de façon générale seulement à une solution quasi-optimale.

La stratégie proposée est itérative. À chaque étape, on recherche parmi les regroupements de deux catégories ligne ou de deux catégories colonne celui qui maximise le critère d'association $\theta(T)$ retenu. Formellement, en notant (P_x^k, P_y^k) la partition obtenue à l'étape k , il s'agit à chaque étape k de chercher la solution (P_x^k, P_y^k) du programme :

$$\begin{cases} \max_{P_x, P_y} \theta(T(P_x, P_y)) \\ \text{s.c. } P_x = P_x^{(k-1)} \text{ et } P_y \in \mathcal{P}_y^{(k-1)} \\ \text{ou} \\ P_x \in \mathcal{P}_x^{(k-1)} \text{ et } P_y = P_y^{(k-1)} \end{cases} \quad (3)$$

où $\mathcal{P}_x^{(k-1)}$ désigne l'ensemble des partitions sur la variable x obtenues par un regroupement de deux classes de la partition $P_x^{(k-1)}$.

Pour des variables ordinales, il convient de remplacer $\mathcal{P}_x^{(k-1)}$ et $\mathcal{P}_y^{(k-1)}$ par les ensembles $\mathcal{A}_x^{(k-1)}$ et $\mathcal{A}_y^{(k-1)}$ de partitions obtenues par le regroupement de deux éléments adjacents.

En partant de $T^0 = T_{\ell \times c}$ le tableau initial associé aux catégories les plus fines des variables x et y , l'heuristique consiste alors à rechercher successivement les tableaux $T^k, k = 1, 2, \dots$ définis par les partitions solution de (3). La procédure se poursuit tant que $\theta(T^k) \geq \theta(T^{(k-1)})$ et est arrêtée dès que cette condition n'est plus vérifiée. En d'autre terme on procède successivement au regroupement de deux catégories qui maximise l'accroissement du critère d'association jusqu'à ce que seul un accroissement négatif puisse être obtenu.

Le *regroupement quasi-optimal* est le couple (P_x^k, P_y^k) solution de (3) à l'étape k où :

$$\theta(T^{(k+1)}) - \theta(T^k) < 0 \quad \text{et} \quad \theta(T^k) - \theta(T^{(k-1)}) \geq 0$$

Par convention, on fixe le critère d'association $\theta(T)$ à zéro pour toute table ayant une seule ligne ou colonne. L'algorithme conduit ainsi à une table 1×1 ne contenant qu'une valeur si et seulement si toutes les lignes, et donc toutes les colonnes, sont identiquement distribuées.

Sur le plan de la complexité, la borne supérieure du nombre de cas explorés par l'algorithme est :

$$\begin{array}{ll} \frac{\ell(\ell^2 - 1) + c(c^2 - 1)}{6} & \text{dans le cas nominal} \\ \frac{\ell(\ell - 1) + c(c - 1)}{2} & \text{dans le cas ordinal} \end{array}$$

Par exemple pour des tableaux carrés $c = \ell = 3, 4, 5, 6$, le nombre de cas explorés est respectivement au plus 8, 20, 40, 70 dans le cas nominal et 6, 12, 20, 30 dans le cas ordinal. Pour $c = 10$, la borne supérieure devient respectivement 330 et 90, valeurs qu'il convient de comparer au nombre de cas à examiner pour obtenir l'optimum global, soit 13 milliards pour le cas nominal et plus d'un million pour le cas ordinal.

5 LES CRITÈRES D'ASSOCIATION

Nous rappelons ici les formules des critères d'association considérés. Pour plus de détails, voir par exemple [17].

$$\begin{array}{ll} \text{Statistique du } \chi^2 \text{ de Pearson} & X^2 = \sum_{i=1}^{\ell} \sum_{j=1}^c \frac{(n_{ij} - n_{i.}n_{.j})^2}{(n_{i.}n_{.j})} \\ \text{Statistique du rapport de vraisemblance} & G^2 = 2 \sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{n_{i.}n_{.j}}\right) \end{array}$$

5.1 MESURES D'ASSOCIATION BASÉES SUR LE χ^2 DE PEARSON

	théorique	empirique
ϕ	$\phi = \sqrt{\frac{\sum_i \sum_j (p_{ij} - p_{i \cdot} p_{\cdot j})^2}{\sum_i p_{i \cdot} \sum_j p_{\cdot j}}}$	$\hat{\phi} = \sqrt{\frac{X^2}{n}}$
Contingence	$c_c = \sqrt{\frac{\phi^2}{1 + \phi^2}}$	$\hat{c}_c = \sqrt{\frac{X^2}{n + X^2}}$
Tschuprow	$t = \sqrt{\frac{\phi^2}{\sqrt{(\ell - 1)(c - 1)}}}$	$\hat{t} = \sqrt{\frac{X^2}{n \sqrt{(\ell - 1)(c - 1)}}}$
Cramer	$v = \sqrt{\frac{\phi^2}{\min\{\ell, c\} - 1}}$	$\hat{v} = \sqrt{\frac{X^2}{n(\min\{\ell, c\} - 1)}}$

5.2 MESURES NOMINALES DE TYPE PRE

 τ de Goodman-Kruskal

$$\tau_{y \leftarrow x} = \frac{\sum_i \sum_j \frac{p_{ij}^2}{p_{i \cdot}} - \sum_j p_{\cdot j}^2}{1 - \sum_j p_{\cdot j}^2} \quad \hat{\tau}_{y \leftarrow x} = \frac{n \sum_i \sum_j \frac{n_{ij}^2}{n_{i \cdot}} - \sum_j n_{\cdot j}^2}{n^2 - \sum_j n_{\cdot j}^2}$$

Coefficient d'incertitude u de Theil

$$u_{y \leftarrow x} = \frac{\sum_i \sum_j p_{ij} \log_2 \left(\frac{p_{i \cdot} p_{\cdot j}}{p_{ij}} \right)}{\sum_j p_{\cdot j} \log_2 p_{\cdot j}} \quad \hat{u}_{y \leftarrow x} = \frac{n \log_2 n - \sum_i \sum_j n_{ij} \log_2 \left(\frac{n_{i \cdot} n_{\cdot j}}{n_{ij}} \right)}{n \log_2 n - \sum_j n_{\cdot j} \log_2 n_{\cdot j}}$$

5.3 MESURES ORDINALES D'ASSOCIATION

On note respectivement π^c , π^d , π_x , π_y la probabilité d'une paire $\{(x_i, y_i), (x_j, y_j)\}$ avec un ordre concordant, i.e. $x_i > x_j$ et $y_i > y_j$, avec un ordre discordant, i.e. $x_i > x_j$ et $y_i < y_j$, avec égalité sur x seulement et avec égalité sur y seulement. De même, on note m^c , m^d , m_x and m_y le nombre de paires d'observations respectivement concordantes, discordantes, avec égalité sur x seulement et avec égalité sur y seulement.

 γ de Goodman-Kruskal

$$\gamma = \frac{\pi^c - \pi^d}{\pi^c + \pi^d} \quad \hat{\gamma} = \frac{m^c - m^d}{m^c + m^d}$$

 d de Somers

$$d_{y \leftarrow x} = \frac{\pi^c - \pi^d}{\pi^c + \pi^d + \pi_y} \quad \hat{d}_{y \leftarrow x} = \frac{m^c - m^d}{m^c + m^d + m_y}$$

τ_b de Kendall

$$\tau_b = \frac{\pi^c - \pi^d}{\sqrt{(\pi^c + \pi^d + \pi_x)(\pi^c + \pi^d + \pi_y)}} \quad \hat{\tau}_b = \frac{m^c - m^d}{\sqrt{(m^c + m^d + m_x)(m^c + m^d + m_y)}}$$

τ_c de Kendall

$$\tau_c = (\pi^c - \pi^d) \left(\frac{\min\{\ell, c\}}{\min\{\ell, c\} - 1} \right) \quad \hat{\tau}_c = \frac{m^c - m^d}{m_{tot}} \left(\frac{\min\{\ell, c\}}{\min\{\ell, c\} - 1} \right)$$

6 REGROUPEMENT DE DEUX CATÉGORIES

Afin d'étudier formellement l'effet d'un regroupement de deux catégories, on explicite analytiquement cet effet sur un choix de critères d'association. On traite tout d'abord le cas des mesures d'associations fondées sur le χ^2 de Pearson, puis les mesures nominales de type PRE et finalement les mesures pour variables ordinales. Pour les mesures symétriques, on considère le regroupement des colonnes j et k . Pour les mesures asymétriques (mesures PRE et d de Somers), on retient y comme variable dépendante et on examine également l'effet d'un regroupement des catégories i et s de la variable indépendante.

L'étude se limite aux versions empiriques des mesures d'association. Les résultats s'étendent cependant aisément aux mesures d'association théoriques exprimées en termes de probabilités.

6.1 VARIATION DES MESURES FONDÉES SUR LE χ^2 DE PEARSON

La variation du X^2 suite à l'agrégation des colonnes j et k est :

$$\Delta_y X^2 = \frac{1}{n} \sum_{i=1}^{\ell} \left(\frac{(n(n_{ij} + n_{ik}) - n_i(n_{.j} + n_{.k}))^2}{n_i(n_{.j} + n_{.k})} - \frac{(nn_{ij} - n_i n_{.j})^2}{n_i n_{.j}} - \frac{(nn_{ik} - n_i n_{.k})^2}{n_i n_{.k}} \right) \quad (4)$$

En développant et en simplifiant le terme sous le signe de sommation cette variation s'écrit :

$$\Delta_y X^2 = \frac{-n}{n_{.j} n_{.k} (n_{.j} + n_{.k})} \sum_i \frac{(n_{.j} n_{ik} - n_{.k} n_{ij})^2}{n_i} \quad (5)$$

Cette quantité est non positive. Un regroupement de catégories ne peut donc en aucun cas augmenter le X^2 . Au mieux, la variation est nulle. Ceci se produit lorsqu'on a l'équivalence distributionnelle des deux colonnes :

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{ik}}{n_{.k}} \Leftrightarrow (n_{.j} n_{ik} - n_{.k} n_{ij})^2 = 0$$

La réduction du X^2 est d'autant plus importante, que l'écart entre les distributions est grand.

Le coefficient de contingence ϕ est une fonction croissante de X^2 et sa variation est donc également non positive.

Le t de Tschuprow et le v de Cramer sont fonctions croissantes de X^2 mais décroissantes du nombre de catégories. Une réduction de c ou ℓ peut donc compenser la réduction de X^2 et conduire à un accroissement de la valeur de ces deux mesures d'association. En particulier, dans le cas de l'agrégation de deux colonnes lorsque $c \leq \ell$, on a :

$$\Delta_y v > 0 \Leftrightarrow \frac{X^2 + \Delta_y X^2}{X^2} > \frac{c-2}{c-1}$$

Pour $c = 3$ par exemple, on a un accroissement du v de Cramer tant que $-\Delta_y X^2$ reste inférieur $X^2/2$.

Le regroupement de deux catégories n'affecte le dénominateur dans l'expression du v de Cramer que si elle porte sur la variable qui a le moins de catégories. Le v de Cramer ne peut donc augmenter que dans ce cas.

6.2 VARIATION DES MESURES DE TYPE PRE

Les mesures de type PRE sont asymétriques par construction. Il convient alors de distinguer le cas du regroupement sur la variable dépendante (y , colonne dans notre cas) et indépendante (x , ligne).

Notons $S_y^{\tau_{y \leftarrow x}} = \sum_j n_{.j}^2$ et $S_{yx}^{\tau_{y \leftarrow x}} = n \sum_i \sum_j n_{ij}^2 / n_{i.}$, où les indices x et y indiquent respectivement que la quantité S est sensible à un regroupement sur la variable x (ligne) ou y (colonne). On a :

$$\tau_{y \leftarrow x} = \frac{S_y^{\tau_{y \leftarrow x}} - S_{yx}^{\tau_{y \leftarrow x}}}{n^2 - S_y^{\tau_{y \leftarrow x}}}$$

d'où il apparaît clairement que la variation $\Delta_x \tau_{y \leftarrow x}$ suite à un regroupement sur la variable indépendante x peut être analysée par le biais de la seule variation $\Delta_x S_{yx}^{\tau_{y \leftarrow x}}$, tandis que pour la variation $\Delta_y \tau_{y \leftarrow x}$ suite à un regroupement sur la variable dépendante y , on doit prendre en compte les variations de $S_y^{\tau_{y \leftarrow x}}$ et $S_{yx}^{\tau_{y \leftarrow x}}$.

Les variations à considérer sont :

$$\Delta_y S_y^{\tau_{y \leftarrow x}} = (n_{.j} + n_{.k})^2 - n_{.j}^2 - n_{.k}^2 \quad (6)$$

$$\Delta_y S_{yx}^{\tau_{y \leftarrow x}} = n \sum_i \frac{(n_{ij} + n_{ik})^2 - n_{ij}^2 - n_{ik}^2}{n_{i.}} \quad (7)$$

$$\Delta_x S_{yx}^{\tau_{y \leftarrow x}} = n \sum_j \left(\frac{(n_{ij} + n_{sj})^2}{n_{i.} + n_{s.}} - \frac{n_{ij}^2}{n_{i.}} - \frac{n_{sj}^2}{n_{s.}} \right) \quad (8)$$

De même, pour le coefficient d'incertitude de Theil, on a, en notant $S_y^{u_{y \leftarrow x}} = \sum_j n_{.j} \log_2 n_{.j}$ et $S_{yx}^{u_{y \leftarrow x}} = \sum_j \sum_i n_{ij} \log_2 (n_{i.} n_{.j} / n_{ij})$:

$$u_{y \leftarrow x} = \frac{n \log_2 n - S_y^{u_{y \leftarrow x}}}{n \log_2 n - S_{yx}^{u_{y \leftarrow x}}}$$

Pour la variation $\Delta_x u_{y \leftarrow x}$ il suffit d'étudier la variation $\Delta_x S_{yx}^{u_{y \leftarrow x}}$ de la double somme du numérateur, tandis que pour la variation suite à un regroupement sur y , on doit prendre en compte les variations de $S_y^{u_{y \leftarrow x}}$ et $S_{yx}^{u_{y \leftarrow x}}$.

Les variations à considérer sont :

$$\Delta_y S_{yx}^{u_{y \leftarrow x}} = (n_{.j} + n_{.k}) \log_2(n_{.j} + n_{.k}) - n_{.j} \log_2 n_{.j} - n_{.k} \log_2 n_{.k} \quad (9)$$

$$\Delta_y S_{yx}^{u_{y \leftarrow x}} = \sum_i \left((n_{ij} + n_{ik}) \log_2 \left(\frac{(n_{.j} + n_{.k}) n_{i.}}{n_{ij} + n_{ik}} \right) - n_{ij} \log_2 \left(\frac{n_{i.} n_{.j}}{n_{ij}} \right) - n_{ik} \log_2 \left(\frac{n_{i.} n_{.k}}{n_{ik}} \right) \right) \quad (10)$$

$$\Delta_{y.} S_{yx}^{u_{y \leftarrow x}} = \sum_j \left((n_{ij} + n_{sj}) \log_2 \left(\frac{n_{.j} (n_{i.} + n_{s.})}{n_{ij} + n_{sj}} \right) - n_{ij} \log_2 \left(\frac{n_{i.} n_{.j}}{n_{ij}} \right) - n_{sj} \log_2 \left(\frac{n_{s.} n_{.j}}{n_{sj}} \right) \right) \quad (11)$$

6.2.1 Regroupement sur la variable indépendante x

Les quantités S_y étant insensibles à un regroupement sur x , les variations $\Delta_x S_y$ sont nulles. Les variations de $\tau_{y \leftarrow x}$ et $u_{y \leftarrow x}$ suite au regroupement de deux lignes (catégories de x) sont alors :

$$\Delta_x \tau_{y \leftarrow x} = \frac{\Delta_x S_{yx}^{\tau_{y \leftarrow x}}}{n^2 - S_y^{\tau_{y \leftarrow x}}} \quad (12)$$

$$\Delta_x u_{y \leftarrow x} = \frac{-\Delta_x S_{yx}^{u_{y \leftarrow x}}}{n \log_2 n - S_y^{u_{y \leftarrow x}}} \quad (13)$$

Les dénominateurs sont dans les deux expressions ci-dessus des quantités positives indépendantes du regroupement. Le maximum de la variation de la mesure correspond alors au maximum de la variation du numérateur, soit de $\Delta_x S_{yx}$.

On peut vérifier que $\Delta_x \tau_{y \leftarrow x}$ que $\Delta_x u_{y \leftarrow x}$ sont non positifs. Ces variations sont nulles lorsque les deux lignes agrégées sont colinéaires, soit lorsque $n_{sj}/n_{s.} = n_{ij}/n_{i.}$.

La variation $\Delta_{y.} S_{yx}^{\tau_{y \leftarrow x}}$ par exemple peut s'écrire sous la forme :

$$\Delta_{y.} S_{yx}^{\tau_{y \leftarrow x}} = n \sum_j \frac{-(n_{ij} n_{s.} - n_{sj} n_{i.})^2}{n_{i.} n_{s.} (n_{i.} + n_{s.})}$$

qui est clairement non positive. Le regroupement de catégories de la variable indépendante ne peut pas alors accroître la valeur de la mesure d'association $\tau_{y \leftarrow x}$. Ceci est en fait assez intuitif : le regroupement ne peut pas accroître le contenu prévisionnel de la variable indépendante.

On établit un résultat similaire pour le coefficient d'incertitude $u_{y \leftarrow x}$. En effet, $\Delta_x S_{yx}^{u_{y \leftarrow x}}$ s'écrit :

$$\Delta_x S_{yx}^{u_{y \leftarrow x}} = n \sum_j \left(n_{ij} \log_2 \left(\frac{n_{ij}}{n_{i.}} \right) + n_{sj} \log_2 \left(\frac{n_{sj}}{n_{s.}} \right) - (n_{ij} + n_{sj}) \log_2 \left(\frac{n_{ij} + n_{sj}}{n_{i.} + n_{s.}} \right) \right) \quad (14)$$

Le terme sous le signe de sommation est non négatif. En effet, il est de la forme :

$$f(a, b, c, d) = a \log_2 \left(\frac{a}{b} \right) + c \log_2 \left(\frac{c}{d} \right) - (a+c) \log_2 \left(\frac{a+c}{b+d} \right)$$

avec $0 \leq a \leq b$ et $0 \leq c \leq d$. La fonction f atteint son minimum en $a/b = c/d = (a+c)/(b+d)$ où l'on a $f(a, b, c, d) = 0$ et ne peut donc pas être négative. Il en est donc de même de (14) ce qui implique une variation (13) de $u_{y \leftarrow x}$ non positive.

6.2.2 Regroupement sur la variable dépendante y

Pour un regroupement sur la variable dépendante (colonne), on doit également tenir compte des changements dans les totaux $n_{.j}$ des colonnes concernées qui impliquent des variations dans les dénominateurs des formules définissant $\tau_{y \leftarrow x}$ et $u_{y \leftarrow x}$. Les variations de ces mesures sont dans ce cas :

$$\Delta_y \tau_{y \leftarrow x} = \frac{(n^2 - S_y^{\tau_{y \leftarrow x}}) \Delta_y S_{yx}^{\tau_{y \leftarrow x}} - (n^2 - S_{yx}^{\tau_{y \leftarrow x}}) \Delta_y S_y^{\tau_{y \leftarrow x}}}{(n^2 - S_y^{\tau_{y \leftarrow x}})^2 - (n^2 - S_{yx}^{\tau_{y \leftarrow x}}) \Delta_y S_y^{\tau_{y \leftarrow x}}} \quad (15)$$

$$\Delta_y u_{y \leftarrow x} = \frac{(n \log_2 n - S_y^{u_{y \leftarrow x}}) \Delta_y S_y^{u_{y \leftarrow x}} - (n \log_2 n - S_{yx}^{u_{y \leftarrow x}}) \Delta_y S_{yx}^{u_{y \leftarrow x}}}{(n \log_2 n - S_y^{u_{y \leftarrow x}})^2 - (n \log_2 n - S_{yx}^{u_{y \leftarrow x}}) \Delta_y S_y^{u_{y \leftarrow x}}} \quad (16)$$

Les dénominateurs des deux expressions sont positifs. Le signe de la variation est donc déterminé pour chacune des deux mesures par celui du numérateur.

En ce qui concerne $\tau_{y \leftarrow x}$, il ressort clairement des formules (6) et (7) que $\Delta_y S_y^{\tau_{y \leftarrow x}}$ et $\Delta_y S_{yx}^{\tau_{y \leftarrow x}}$ sont tous les deux positifs. On peut donc avoir aussi bien des variations positives que négatives :

$$\Delta_y \tau_{y \leftarrow x} \geq 0 \Leftrightarrow \frac{\Delta_y S_{yx}^{\tau_{y \leftarrow x}}}{\Delta_y S_y^{\tau_{y \leftarrow x}}} \geq \frac{(n^2 - S_{yx}^{\tau_{y \leftarrow x}})}{(n^2 - S_y^{\tau_{y \leftarrow x}})} \quad (17)$$

$$\Delta_y \tau_{y \leftarrow x} \leq 0 \Leftrightarrow \frac{\Delta_y S_{yx}^{\tau_{y \leftarrow x}}}{\Delta_y S_y^{\tau_{y \leftarrow x}}} \leq \frac{(n^2 - S_{yx}^{\tau_{y \leftarrow x}})}{(n^2 - S_y^{\tau_{y \leftarrow x}})} \quad (18)$$

Pour u_{xy} , il ressort également des formules (9) et (10) que $\Delta_y S_y^{u_{y \leftarrow x}}$ et $\Delta_y S_{yx}^{u_{y \leftarrow x}}$ sont tous deux positifs. La variation peut donc là aussi être positive ou négative :

$$\Delta_y u_{y \leftarrow x} \geq 0 \Leftrightarrow \frac{\Delta_y S_y^{u_{y \leftarrow x}}}{\Delta_y S_{yx}^{u_{y \leftarrow x}}} \geq \frac{(n^2 - S_y^{u_{y \leftarrow x}})}{(n^2 - S_{yx}^{u_{y \leftarrow x}})} \quad (19)$$

$$\Delta_y u_{y \leftarrow x} \leq 0 \Leftrightarrow \frac{\Delta_y S_y^{u_{y \leftarrow x}}}{\Delta_y S_{yx}^{u_{y \leftarrow x}}} \leq \frac{(n^2 - S_y^{u_{y \leftarrow x}})}{(n^2 - S_{yx}^{u_{y \leftarrow x}})} \quad (20)$$

6.3 MESURES ORDINALES

On considère ici les mesures pour variables ordinales fondées sur les notions de paires concordantes et discordantes, deux observations étant dites concordantes si on a la même relation d'ordre sur les variables x et y , et discordantes si la relation d'ordre est inversée. L'association entre variables catégorielles pouvant être négative, il convient de maximiser la valeur absolue des mesures. Par ailleurs, on se limite à l'étude de l'effet du regroupement de catégories adjacentes.

Les mesures examinées sont le γ de Goodman et Kruskal, le τ_b de Kendall, le τ_c de Kendall et Stuart et les d_y et d_x asymétriques de Somers.

Il s'agit donc d'étudier les quantités :

$$\begin{aligned}
 m^c &= \sum_{i=1}^{\ell-1} \sum_{j=1}^{c-1} \left(n_{ij} \sum_{i'>i} \sum_{j'>j} n_{i'j'} \right) \\
 m^d &= \sum_{i=1}^{\ell-1} \sum_{j=2}^c \left(n_{ij} \sum_{i'>i} \sum_{j'<j} n_{i'j'} \right) \\
 m_x &= \sum_{i=1}^{\ell} \sum_{j=1}^{c-1} \left(n_{ij} \sum_{j'>j} n_{ij'} \right) \\
 m_y &= \sum_{j=1}^c \sum_{i=1}^{\ell-1} \left(n_{ij} \sum_{i'>i} n_{i'j} \right) \\
 m_{xy} &= \sum_{i=1}^c \sum_{j=1}^c \frac{n_{ij}(n_{ij} - 1)}{2} \\
 m_{tot} &= m^c + m^d + m_x + m_y + m_{xy} = \frac{n(n-1)}{2}
 \end{aligned}$$

dont les variations $\Delta_y m$ suite à l'agrégation des colonnes k et $k+1$ sont :

$$\Delta_y m^c = - \sum_{i=1}^{\ell-1} n_{ik} \sum_{i'>i} n_{i'(k+1)} \quad (21)$$

$$\Delta_y m^d = - \sum_{i=1}^{\ell-1} n_{i(k+1)} \sum_{i'>i} n_{i'k} \quad (22)$$

$$\Delta_y m_x = - \sum_{i=1}^{\ell} n_{ik} n_{i(k+1)} \quad (23)$$

$$\Delta_y m_y = \sum_{i=1}^{\ell-1} \left(n_{ik} \sum_{i'>i} n_{i'(k+1)} + n_{i(k+1)} \sum_{i'>i} n_{i'k} \right) \quad (24)$$

$$\Delta_y m_{xy} = \sum_{i=1}^{\ell} n_{ik} n_{i(k+1)} \quad (25)$$

Le nombre total m_{tot} de paires restant évidemment inchangé.

Le regroupement de catégories de y transforme évidemment des inégalités sur cette variable en égalités. Ceci se traduit par un transfert de paires comptabilisées dans m^c et m^d vers m_y et de m_x vers m_{xy} .

Discussion

En cas d'équivalence distributionnelle ($n_{i(k+1)} = \alpha n_{ik}$) des deux colonnes k et $k + 1$, la réduction de m^c et de m^d est identiques. La différence $m^c - m^d$, numérateur commun de toutes les mesures, reste donc constante tandis que la somme $m^c + m^d$ diminue. Le regroupement de colonnes identiquement distribuées induit dès lors un accroissement du γ de Goodman et Kruskal. Pour le d_y de Somers, pertinent lorsque y est la variable dépendante, la diminution de $m^c + m^d$ est compensée par l'accroissement de m_y , et la mesure reste donc insensible à l'agrégation de colonnes identiquement distribuées. Pour d_x , c'est-à-dire lorsqu'on agrège des catégories de la variable indépendante, la diminution de $m^c + m^d$ est amplifiée par celle de m_x et l'on observe alors un renforcement de l'association. En ce qui concerne le τ_b , le dénominateur étant la moyenne géométrique de celui de d_y et de d_x , on a également un accroissement de l'association, mais moins marqué que pour le d_y . Enfin, le τ_c est affecté par un regroupement de catégories identiquement distribuées uniquement par l'effet sur $\min\{\ell, c\}$. Le τ_c augmente lorsque $c \leq \ell$ et reste inchangé sinon.

En dehors de l'équivalence distributionnelle, toutes les situations peuvent se présenter. Si dans la colonne k les n_{ik} ont tendance à être plus importants pour les petits i que pour les grands, et dans la colonne $k + 1$ les $n_{i(k+1)}$ ont tendance à être plus importants pour les grands i que pour les petits, la réduction de m^c sera plus importante que celle de m^d . Le regroupement devrait alors diminuer l'association si elle est positive et la renforcer si elle est négative.

7 FIABILITÉ DE L'HEURISTIQUE

Afin d'apprécier la fiabilité de l'heuristique proposée au paragraphe 4.2, nous rapportons ici quelques résultats empiriques de simulations pour des tables de contingence de taille 4×4 et 6×6 . Dans chacun de ces deux cas, 500 tables ont été générées de façon aléatoire. Pour chaque tableau généré, les valeurs du t de Tschuprow ou du v de Cramer correspondant à l'agrégation quasi-optimale et à l'optimum global ont été répertoriées et comparées. Seul le cas de catégories non ordonnées a été étudié. Le Tableau 2 donne les caractéristiques des différences relevées et la Figure 1 montre comment les écarts se distribuent selon la valeur optimale de la mesure d'association pour les cas des tableaux de départ de taille 6×6 .

Comme les écarts sont bornés supérieurement par la valeur de l'optimum global, il n'est pas étonnant d'observer une augmentation de la dispersion avec la valeur de cet optimum. Si le pourcentage d'optima globaux manqués par l'heuristique est relativement important (environ 50 %), la sous-évaluation que donne l'heuristique reste petit. Il convient aussi de préciser que la distribution des écarts est fortement asymétrique avec un coefficient d'asymétrie pour les écarts non nuls supérieur à 1,5 pour le v de Cramer et

à 1 pour le t de Tschuprow. On peut donc s'attendre à ce que l'écart soit en général inférieur à l'écart moyen donné.

Tableau 2. Simulations : différences entre quasi-optimum et optimum global (nominal)

Tableaux de départ	Tschuprow		Cramer	
	4 × 4	6 × 6	4 × 4	6 × 6
Cas simulés	500	500	500	500
Optima non atteints	125	261	157	263
en pourcent	25.0 %	52.2 %	31.4 %	52.6 %
Écarts non nuls				
moyenne	0.030	0.027	0.010	0.015
écart type	0.028	0.027	0.012	0.016
asymétrie	1.04	1.13	1.73	1.50
Y compris écarts nuls				
moyenne	0.007	0.014	0.003	0.008
écart type	0.019	0.024	0.009	0.014
asymétrie	3.06	1.97	3.50	2.26
Différence maximale	0.106	0.138	0.063	0.100
Écart relatifs non nuls				
moyenne	1.6 %	3.3 %	2.3 %	3.1 %
maximum	16.3 %	24.1 %	13.8 %	18.0 %

Les erreurs réalisées sont évidemment à mettre en regard du gain de performance que fournit l'heuristique. Pour en donner une idée plus précise, on peut indiquer que la recherche des optima globaux de 500 tables 6×6 requiert environ 20 heures sur un micro PII 300, tandis que les quasi-optima sont obtenus en quelques secondes.

8 CONCLUSION

Cet article est une contribution au problème de la recherche des partitions des catégories ligne et colonne qui maximisent l'association. Dans ce contexte, les résultats présentés ici, que ce soit sur la complexité de la solution optimale, sur la sensibilité des critères d'association à l'agrégation de catégories ou encore l'heuristique proposée ne constituent

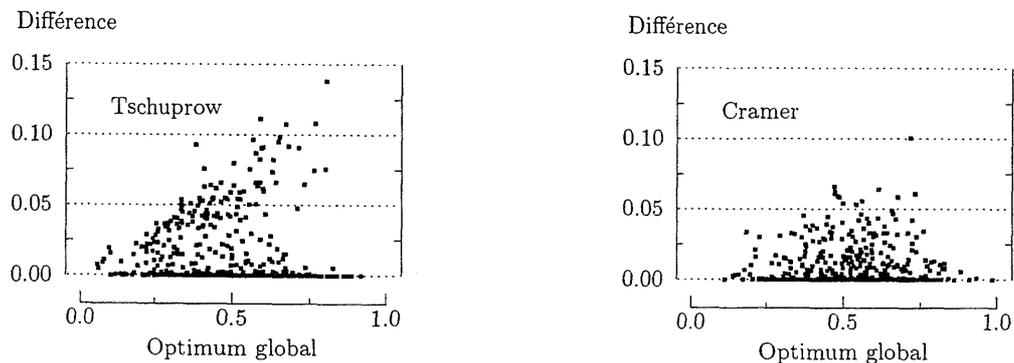


Figure 1. Différences selon optimum global, 500 tables 6×6 (cas nominal)

qu'une étape préliminaire. Il reste beaucoup à faire. D'une part il s'agit d'étudier plus finement les propriétés de l'algorithme introduit au point 4.2, en particulier pour le cas de catégories ordonnées pertinent pour les problèmes de discrétisation. Il est également prévu d'explorer une procédure hiérarchique descendante, où l'on procéderait par dichotomisations successives à partir d'un tableau totalement agrégé. Nous avons également commencé à réfléchir à la possibilité de procéder à une partition optimale directe de la table de contingence, un peu dans l'esprit des diagrammes de Bertin [5], et d'en déduire ensuite l'agrégation correspondante des marges. Enfin, pour mieux tenir compte de l'instabilité en généralisation de résultats fondés sur de petits effectifs, il est prévu d'introduire dans l'algorithme les estimations de Laplace $\hat{p}_{ij}^\lambda = \frac{n_{ij} + \lambda}{n + rc\lambda}$ des probabilités. L'utilisation de ces estimateurs permet en effet d'accroître la robustesse de la solution avec la valeur de λ .

Remerciements : Les auteurs tiennent à remercier deux rapporteurs anonymes dont les suggestions ont contribué à améliorer sensiblement la qualité de cet article.

BIBLIOGRAPHIE

- [1] ANDERBERG, M., *Cluster Analysis for Application*, New York, Academic Press, 1973.
- [2] AURAY, J.-P., DURU, G., ZIGHED, D.A., *Analyse des données multidimensionnelles*, éditions A. Lacassagne, Lyon, 1990.
- [3] BELL, E.T., « The iterated exponential numbers », *Ann. Math.* 39, 1938, p. 539-557.
- [4] BENZÉCRI, J.-P., *Analyse des données. Tome 2: Analyse des correspondances*, Paris, Dunod, 1973.
- [5] BERTIN, J., *La graphique et le traitement graphique de l'information*, Paris, Flammarion, 1977.
- [6] BLOCK, H., « Simultaneous clustering of objects and variables », *Data Analysis and Informatics*, Diday et al. eds., 1979, p. 187-203.
- [7] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J., *Classification And Regression Trees*, New York, Chapman and Hall, 1993.
- [8] CELEUX, G., DIDAY, E., GOVAERT, G., LECHEVALLIER, Y., RALAMBONDRAINY, H., *Classification automatique des données. Informatique*, Paris, Dunod, 1988.
- [9] FISHER, W. D., « Optimal aggregation in multi-equation prediction models », *Econometrica* 30, 1962, p. 744-769.
- [10] FISHER, W. D., *Clustering and Aggregation in Economics*, The John Hopkins Press, Baltimore, 1969.
- [11] GILULA, Z., KRIEGER, A. M., « The decomposability and monotonicity of Pearson's chi-square for collapsed contingency tables with applications », *Journal of the American Statistical Association* 78, 1983, p. 176-180.
- [12] GOVAERT, G., « Classification simultanée de tableaux binaires », *Data Analysis and Informatics* 3, Diday et al. eds., Amsterdam, North-Holland, 1984, p. 223-236.

- [13] GOVAERT, G., « Simultaneous clustering of rows and columns », *Control and Cybernetics* 24(4), 1995, p. 438-458.
- [14] GREENACRE, M., « Clustering the rows and columns of a contingency table », *Journal of Classification* 5, 1988, p. 39-51.
- [15] HIROTSU, C., « Defining the pattern of association in two-way contingency tables », *Biometrika* 70, 1983, p. 579-589.
- [16] JOBSON, J. D., *Applied Multivariate Data Analysis*, volume II: Categorical and Multivariate Methods, New York, Springer-Verlag, 1992.
- [17] OLSZAK, M., RITSCHARD, G., « The behaviour of nominal and ordinal partial association measures », *The Statistician* 44(2), 1995, p. 195-212.
- [18] RAKOTOMALALA, R., ZIGHED, D. A., « Mesures PRE dans les graphes d'induction: une approche statistique de l'arbitrage généralité-précision », *Apprentissage: des principes naturels aux méthodes artificielles*, Ritschard G., Berchtold A., Duc F., Zighed D. A., eds., Paris, Hermes Science Publications, 1998, p. 37-60.
- [19] ZIGHED, D. A., RAKOTOMALALA, R., *Graphes d'induction: apprentissage et data mining*, Paris, Hermes Science Publications, 2000.