



Généralisation de l'analyse des correspondances à la comparaison de tableaux de fréquence

Brigitte Escofier

► To cite this version:

Brigitte Escofier. Généralisation de l'analyse des correspondances à la comparaison de tableaux de fréquence. [Rapport de recherche] RR-0207, INRIA. 1983. inria-00076351

HAL Id: inria-00076351

<https://inria.hal.science/inria-00076351>

Submitted on 24 May 2006

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.



CENTRE DE RENNES
IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Volveau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

Rapports de Recherche

N° 207

**GÉNÉRALISATION DE L'ANALYSE
DES CORRESPONDANCES
À LA COMPARAISON
DE TABLEAUX DE FRÉQUENCE**

Brigitte ESCOFIER

Mai 1983

GENERALISATION DE L'ANALYSE DES CORRESPONDANCES
A LA COMPARAISON DE TABLEAUX DE FREQUENCE

Brigitte ESCOPIER

Publication Interne n° 190
Mars 1983
35 pages

RESUME : Nous proposons ici une méthode de comparaison de tableaux de fréquence définis sur le même couple d'ensembles I et J. Cette méthode a pour but de mettre en évidence systématiquement et d'analyser les distances entre les mêmes lignes des différents tableaux. Elle est dérivée de l'analyse des correspondances et permet une représentation simultanée de toutes les lignes de tous les tableaux et de l'ensemble J des colonnes. Des formes de type barycentrique lient ces deux représentations et permettent d'expliquer les résultats.

On peut voir aussi cette méthode comme une analyse des différences entre les tableaux, après l'élimination de l'écart entre leur marge sur I, écarts qui peuvent masquer des phénomènes plus complexes.

SUMMARY : In order to compare contingency tables, we propose a generalization of the correspondance analysis. It allows to point out and analyze the differences, even very low, between the same rows of different tables. A simultaneous representation of the columns is used to interpret the results.



GENERALISATION DE L'ANALYSE DES CORRESPONDANCES
A LA COMPARAISON DE TABLEAUX DE FREQUENCE

SOMMAIRE

1. Introduction
2. Notations
3. Quelques méthodes de comparaison de tableaux de fréquence.
 - 3.1 la juxtaposition de tableaux
 - 3.2 la double juxtaposition
 - 3.3 A.C.P. des différences entre profils
 - 3.4 analyse factorielle sphérique
4. Propriétés générales de l'analyse proposée
5. Maximisation de l'inertie intra de $N(I \times T)$
 - 5.1 le principe
 - 5.2 nuage $N(I \times T)$ et nuage recentré $N'(I \times T)$
 - 5.3 calcul des facteurs F_s et F'_s
6. Le nuage $N(J)$
 - 6.1 les coordonnées des points j
 - 6.2 nuage des profils du tableau somme
 - 6.3 $N(J)$ est un nuage recentré
 - 6.4 interprétation des distances dans $N(J)$
 - 6.5 interprétation en termes de correspondance ternaire
7. Ecart entre les marges sur I des tableaux
 - 7.1 le problème de l'écart entre les marges
 - 7.2 le modèle
 - 7.3 analyse de la différence entre données et modèle
8. Représentation simultanée et interprétation des résultats
 - 8.1 facteurs sur $I \times T$: F_s et F'_s
 - 8.2 contribution absolue et qualité de représentation des (i, t)
 - 8.3 facteurs sur J : G_s
 - 8.4 transition de G_s vers F_s et F'_s
 - 8.5 transition de F_s et F'_s vers G_s
 - 8.6 éléments supplémentaires
9. Les deux analyses des lignes et des colonnes
10. Exemple
 - 10.1 les données
 - 10.2 comparaison des profils d'emploi à diplôme égal
 - 10.3 comparaison des profils de diplômes à emploi égal
11. Conclusion

GENERALISATION DE L'ANALYSE DES CORRESPONDANCES

A LA COMPARAISON DE TABLEAUX DE FREQUENCE

1. INTRODUCTION

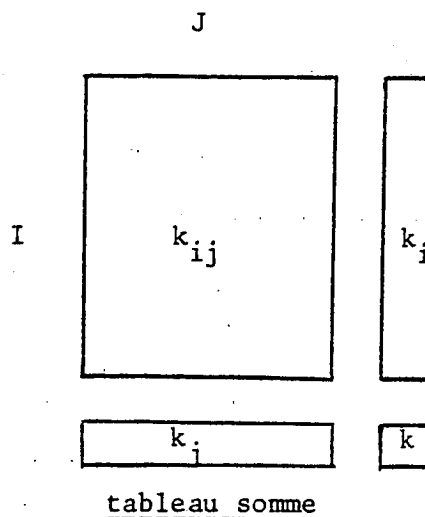
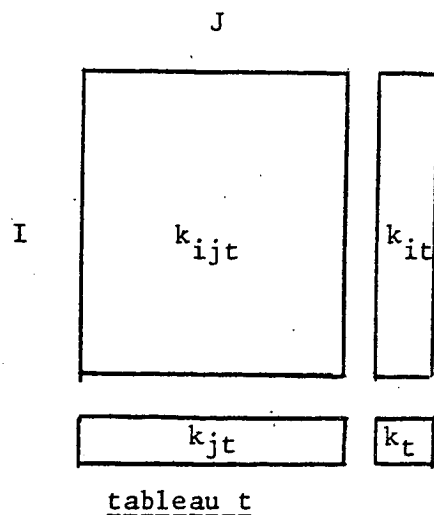
Nous proposons ici une méthode de comparaison de tableaux de fréquence définis sur le même couple d'ensembles I et J. Cette méthode a pour but de mettre en évidence systématiquement et d'analyser les distances entre les mêmes lignes des différents tableaux. Elle est dérivée de l'analyse des correspondances et permet une représentation simultanée de toutes les lignes de tous les tableaux et de l'ensemble J des colonnes. Des formules de type barycentrique lient ces deux représentations et permettent d'expliquer les résultats.

On peut voir aussi cette méthode comme une analyse des différences entre les tableaux, après l'élimination de l'écart entre leur marge sur I, écarts qui peuvent masquer des phénomènes plus complexes.

2. NOTATIONS

Nous notons I l'ensemble des lignes des tableaux et J l'ensemble de leurs colonnes. Les tableaux sont indicés par les éléments d'un ensemble T (évoquant le temps, qui apparaît fréquemment dans ce type de données).

Le terme général du tableau t est noté k_{ijt} , ses marges sur I et sur J sont notées k_{it} et k_{jt} , et son effectif total k_t . Nous considérons aussi la somme de tous les tableaux de fréquence, que nous appelons aussi tableau moyen. Le terme général de cette somme est noté k_{ij} , ses marges k_i et k_j , et son effectif total k.



$$k_{ij} = \sum_t k_{ijt}$$

$$k_{it} = \sum_j k_{ijt}$$

$$k_{jt} = \sum_i k_{ijt}$$

$$k_i = \sum_j k_{ij} = \sum_{j,t} k_{ijt}$$

$$k_j = \sum_i k_{ij} = \sum_{i,t} k_{ijt}$$

$$k_t = \sum_{i,j} k_{ijt}$$

$$k = \sum_{i,j,t} k_{ijt}$$

3. QUELQUES METHODES DE COMPARAISON DE TABLEAUX DE FREQUENCE

Avant d'exposer cette nouvelle méthode, nous rappelons dans ce paragraphe, quelques techniques permettant d'analyser les différences entre plusieurs tableaux de fréquence. Aucune ne permet d'étudier systématiquement les différences entre plus de deux tableaux de fréquence.

3.1. la juxtaposition des tableaux (cf.3)

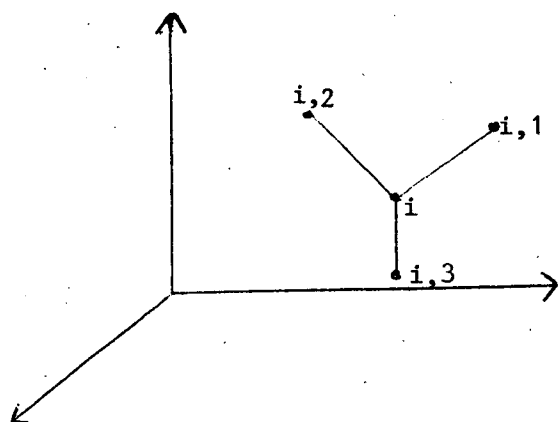
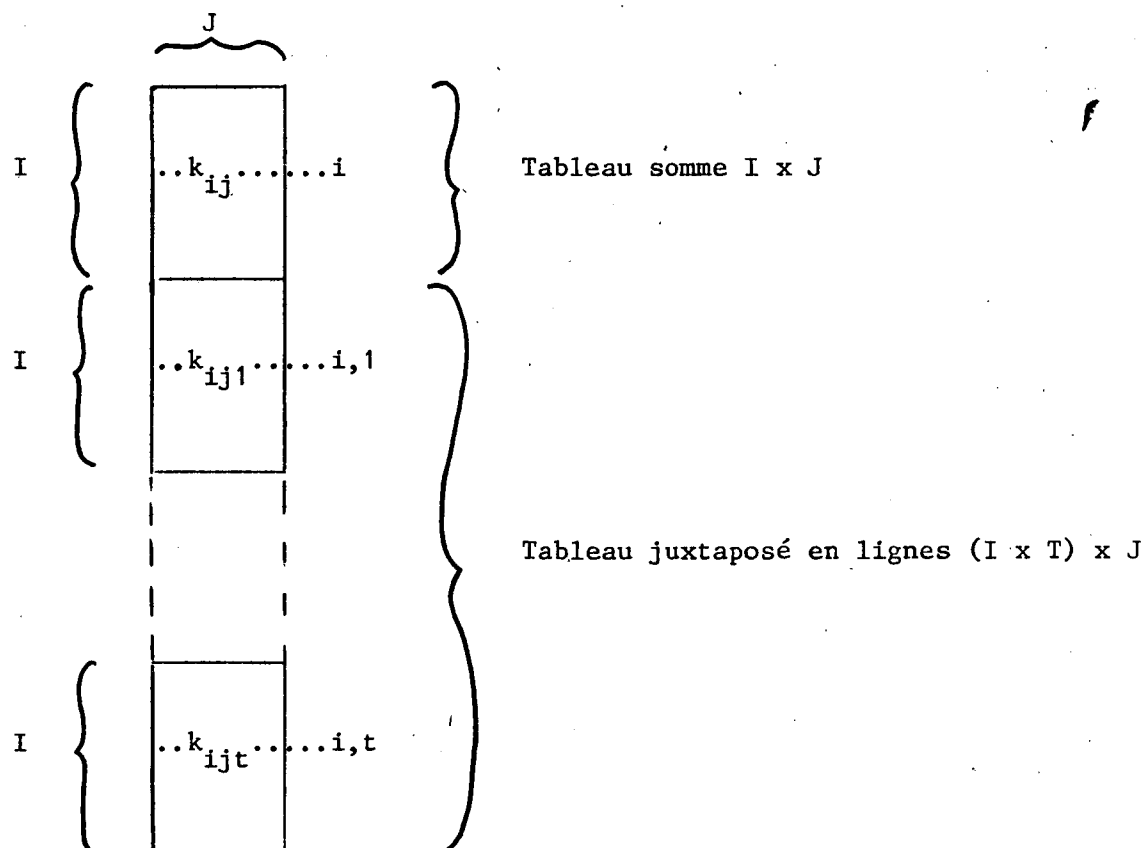
Une technique simple de comparaison consiste à juxtaposer tous les tableaux de fréquence et leur somme et de faire l'analyse des correspondances du grand tableau ainsi obtenu.

Cette technique permet de comparer les profils des lignes des tableaux si on les juxtapose en lignes et les profils de leurs colonnes si on les juxtapose en colonnes. (Nous insistons sur cette méthode car nous nous placerons exactement dans le même cadre).

Considérons, par exemple, les tableaux juxtaposés en lignes. Dans

l'espace $\mathbb{R}_J$ où sont placés les profils des lignes, le profil k_{iJ}/k_i de la ligne i du tableau somme est situé au centre de gravité des profils k_{iJt}/k_{it} des lignes homologues (i,t) de tous les tableaux (le poids affecté à (i,t) est, bien entendu, son poids naturel k_{it}/k).

$$k_{iJ}/k_i = \sum_{t \in T} (k_{iJt}/k_{it}) (k_{it}/k_i)$$



Espace $\mathbb{R}_J$: i est au centre de gravité des (i,t)

L'inertie du nuage $N(IxT)$ des profils des lignes de tous les tableaux peut se décomposer d'après le principe de Huyghens en somme de l'inertie "inter" (Inertie de $N(I)$) et de l'inertie "intra" (Inertie des nuages $N(ixT)$ autour de leur centre de gravité i).

3.1.1 Analyse de $I \times J$ avec $(IxT) \times J$ en supplémentaire

Généralement, on fait l'analyse des correspondances du tableau ci-dessus avec les lignes du tableau somme $I \times J$ en éléments principaux et celles des tableaux $(I, t) \times J$ en éléments supplémentaires.

Ceci revient à projeter le nuage $N(IxT)$ sur les axes principaux d'inertie de $N(I)$, c'est-à-dire, des axes rendant maximum son inertie "inter".

Il est clair que par cette méthode, on extrait des facteurs "moyens" pour tous les tableaux et que les écarts entre les profils d'une même ligne ne sont analysés que sur ces tendances moyennes.

Si les profils homologues des différents tableaux se ressemblent beaucoup plus entre eux que les profils du tableau moyen, leurs différences ne seront pas visibles sur les graphiques. Dans le cas extrême où les nuages $N(ixT)$ des profils d'une même ligne autour de leur centre de gravité sont orthogonaux au nuage $N(I)$ de leurs centres de gravité, leur dispersion, même importante, est totalement ignorée par cette analyse.

3.1.2 Le tableau somme en supplémentaire

On peut aussi faire l'analyse du même tableau avec les lignes du tableau $(IxT) \times J$ en principales et celles du tableau somme $I \times J$ en supplémentaires.

Cette analyse traduira mieux que la précédente les différences entre les lignes homologues puisque les nuages sont projetés sur des axes qui rendent maximum l'inertie totale du nuage $N(IxT)$. L'inertie intra rentre donc en ligne de compte et dans la situation extrême d'orthogonalité évoquée ci-dessus, les différences entre lignes homologues apparaîtront.

Mais, si l'inertie intra (différence entre profils homologues) est beaucoup plus faible que l'inertie inter (différence entre les profils de la somme), les axes obtenus seront proches des axes inter et les différences

entre profils homologues ne seront guère plus visibles que dans l'analyse de $N(I)$.

Cette technique ne permet donc pas de mettre en évidence systématiquement et d'analyser les écarts entre les profils des lignes homologues. En posant le problème dans ce cadre géométrique et dans ces termes, la solution que nous proposons vient facilement à l'esprit : nous analyserons les différences entre les profils des lignes homologues en projetant leur nuage sur des axes rendant maximum l'inertie intra.

3.2. La double juxtaposition

Une technique de comparaison utilisant encore un programme classique d'analyse des correspondances a été proposée par E. Yagolnitzer (cf7).

Elle ne s'applique qu'à la comparaison de deux tableaux seulement et consiste à faire l'analyse d'un grand tableau où les deux tableaux apparaissent deux fois :

$$\begin{matrix} & & I & & J \\ I & & \left(\begin{array}{cc} k_{ij1} & k_{ij2} \\ k_{ij2} & k_{ij1} \end{array} \right) \end{matrix}$$

Dans l'article cité, on montre que l'on obtient deux séries de facteurs. D'une part, ceux de la somme des deux tableaux où les lignes (resp. les colonnes) homologues du tableau analysé sont confondues. D'autre part, des facteurs où les lignes (resp. colonnes) homologues sont opposées. Ces derniers sont les projections des deux nuages des profils des différences $(k_{ij1} - k_{ij2})$ et $(k_{ij2} - k_{ij1})$, profils calculés avec les marges k_i et k_j de la somme. Un exemple d'application est donné dans le même cahier (cf8). Les facteurs de la deuxième série qui sont ceux qui nous intéressent ici, puisqu'ils analysent les différences, sont presque entièrement expliqués par les écarts entre les marges des deux tableaux. Ceci est naturel puisque c'est la différence $(k_{ij1} - k_{ij2})$ qui est prise en compte (et non la différence entre les profils) et lorsque les marges des deux tableaux sont très différentes, cet effet devient prédominant dans l'écart $k_{ij1} - k_{ij2}$.

3.3. A.C.P. des différences entre profils

Cet écueil est évité par P. Clapier et N. Tabard (cf 4) qui font une analyse en composantes principales de la différence entre les profils des lignes des deux tableaux.

Cette méthode se montre efficace mais ne se généralise pas non plus à plus de deux tableaux. D'autre part, les distances considérées dans l'analyse en composantes principales sont, pour des tableaux de fréquence, moins naturelles que la distance du χ^2 .

Remarquons que l'étude des profils des lignes et des profils des colonnes nécessite deux analyses différentes.

3.4. L'analyse factorielle sphérique

Cette méthode proposée par M. Volle et D. Domenges (cf 5) a été conçue, en particulier, pour mettre en évidence les différences entre deux tableaux de fréquence.

On définit d'abord une distance entre ces deux tableaux par la formule suivante :

$$D^2(\text{Tableau 1, Tableau 2}) = \sum_{i,j} (\sqrt{k_{ij1}/k_1} - \sqrt{k_{ij2}/k_2})^2$$

(k_1 et k_2 sont les effectifs des deux tableaux).

Puis, on construit deux nuages qui représentent l'un I dans l'espace $\mathbb{R}^J$, et l'autre J dans l'espace $\mathbb{R}^I$. Les coordonnées des points du premier sont :

$$i \rightsquigarrow \sqrt{k_{iJ1}/k_1} - \sqrt{k_{iJ2}/k_2}$$

Le deuxième nuage se définit d'une manière analogue, il est clair que les inerties de ces deux nuages sont égales entre elles et égales à D^2 . La projection de ces nuages sur leurs axes principaux d'inertie permet d'analyser et de décomposer D^2 . Des formules de transition lient ces deux projections, ce qui permet une représentation simultanée de I et de J.

Cette méthode comme celle de Yagolnizter ne permet pas de comparer plus de deux tableaux. D'autre part, les résultats sont difficiles à interpréter car les distances entre deux points i et i' (ou j et j') ne sont pas les distances de χ^2 et les formules de transition ne sont pas des formules barycentriques. Enfin, l'effet de l'écart entre les marges des deux tableaux risque encore d'être prépondérant puisqu'il n'est pas éliminé dans la différence considérée.

4. PROPRIETES GENERALES DE L'ANALYSE PROPOSEE

La méthode que nous proposons est une analyse systématique des différences entre les profils des lignes de deux ou plusieurs tableaux de fréquence. Nous l'appelons "analyse intra", car elle est basée sur une projection sur des axes rendant maximum l'inertie des profils des lignes homologues autour de leur moyenne. Ainsi ces différences seront toujours mises en évidence, même si elles sont extrêmement faibles relativement aux écarts entre profils d'un même tableau. C'est dans cette optique que la méthode est présentée au paragraphe 5 dans le cadre géométrique précisé au §3.1 dans l'étude des tableaux juxtaposés. Une analyse des colonnes J associée à l'analyse intra des lignes est étudiée au paragraphe 6.

Au paragraphe 7 nous proposons une approche complètement différente de cette méthode. Nous avons vu que certaines techniques de comparaison mettaient essentiellement en évidence les écarts entre les marges des tableaux. Ces écarts peuvent présenter un intérêt et des techniques simples permettent de les analyser (cf 6) ; mais ce n'est qu'un aspect très partiel de la différence entre tableaux. Dans certains cas, les écarts entre l'une des marges de tous les tableaux, peuvent être un artefact qu'il est nécessaire d'éliminer pour qu'il ne masque pas les différences qui nous intéressent réellement. L'analyse intra permet de comparer les tableaux en éliminant l'influence des écarts entre leur marge sur I .

Le paragraphe 8 résume les formules et précise l'interprétation des résultats. Cette interprétation est assez proche de celle d'une analyse des correspondances classiques : les distances entre éléments s'expriment en termes de distance du χ^2 ; les qualités de représentation et les contributions à l'inertie des lignes et des colonnes sont calculées ; des formules de "transition" relativement simples permettent d'expliquer la position des lignes par rapport à celles des colonnes et réciproquement. Les lignes et les colonnes ne jouent pas un rôle symétrique ; au paragraphe 9 on compare les diverses analyses possibles.

Au paragraphe 10, on montre sur un exemple très simple et uniquement pédagogique l'intérêt et l'interprétation des résultats.

5. MAXIMISATION DE L'INERTIE INTRA DE $N(I \times T)$

5.1. Le principe

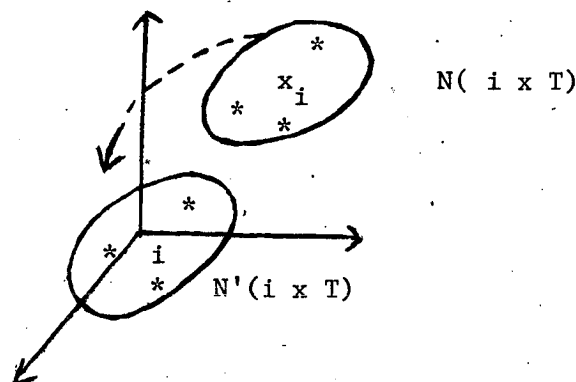
Nous avons déjà indiqué au paragraphe 3.1. l'idée principale de la solution que nous proposons. Nous avons décomposé l'inertie totale du nuage $N(I \times T)$ des lignes des tableaux juxtaposés en inertie inter et inertie intra. La première est l'inertie des centres de gravité i des lignes homologues des différents tableaux. La seconde est l'inertie des lignes homologues autour de leur centre de gravité.

Pour mettre en évidence les différences entre les profils des lignes homologues de tous les tableaux, nous allons projeter le nuage $N(I \times T)$ sur des axes tels que l'inertie intra de sa projection soit maximum. Nous appellerons ces axes, axes d'inertie intra de $N(I \times T)$.

Il est bien clair que la métrique de l'espace $\mathbb{R}_J$ reste la métrique diagonale k/k_j qui est induite à la fois par le tableau $(I \times T) \times J$ et par le tableau somme $I \times J$.

5.2 Nuage $N(I \times T)$ et nuage recentré $N'(I \times T)$

Dans ce paragraphe, nous considérons un nouveau nuage que nous notons $N'(I \times T)$. Ce nuage est déduit de $N(I \times T)$ en recentrant à l'origine chacun des sous nuages $N(i \times T)$.



Espace $\mathbb{R}_J$

Les coordonnées des points (i,t) et i dans les deux nuages $N(I \times T)$ et $N'(I \times T)$ sont résumées dans le tableau ci-dessous :

$N(I \times T)$		$N'(I \times T)$
i	k_{iJ}/k_i	0
i,t	k_{iJt}/k_{it}	$k_{iJt}/k_{it} - k_{iJ}/k_i$

Dans le nuage recentré $N'(I \times T)$ la distance entre deux points d'un même nuage $N(i \times T)$, (i.e. deux lignes homologues dans deux tableaux différents) reste la distance du χ^2 comme dans $N(I \times T)$. Par contre la distance entre deux lignes du tableau somme est nulle et la distance entre deux lignes non homologues n'est plus la distance du χ^2 .

Il est clair que l'inertie totale du nuage recentré $N'(I \times T)$ est égale à l'inertie intra de $N(I \times T)$, dans toutes les directions de R_J . Les axes d'inertie totaux du premier sont donc confondus avec les axes d'inertie intra du second.

Notons F_s et F'_s les coordonnées des projections de $N(I \times T)$ et $N'(I \times T)$ sur ces axes. Nous les appelons respectivement facteurs intra de $N(I \times T)$ et facteurs de $N'(I \times T)$. Entre F_s et F'_s , on a les relations suivantes :

$$F'_s(i) = 0$$

$$F'_s(i,t) = F_s(i,t) - F_s(i)$$

Nous calculerons et représenterons ces deux nuages qui présentent chacun leur intérêt.

Celui du premier est de représenter exactement le nuage étudié dans la juxtaposition de tableau avec les distances du χ^2 classique entre profils.

Celui du second est de représenter directement les écarts des profils (i,t) à leur moyenne (i) et de mettre en évidence, loin de l'origine ceux qui s'écartent le plus de leur moyenne.

5.3. Calcul des facteurs F_s et F'_s

Nous indiquons deux techniques de calcul des facteurs. La seconde est basée sur les propriétés de l'inertie intra et demande l'écriture d'un

programme très proche de celui de l'analyse des correspondances. La première dont l'idée est due à J.P. Benzecri permet d'utiliser un programme classique d'analyse des correspondances en construisant un nouveau tableau.

5.3.1. première technique

Considérons un nouveau tableau croisant $(I \times T) \times J$, noté $\tilde{k}_{(I \times T) \times J}$ dont le terme général est :

$$\tilde{k}_{ijt} = k_{ijt} - k_{ij} k_{it} / k_i + k_j k_{it} / k$$

Ce tableau a les mêmes marges que le tableau juxtaposé $k_{(I \times T) \times J}$:

$$\tilde{k}_{it} = \sum_j \tilde{k}_{ijt} = k_{it}$$

$$\tilde{k}_j = \sum_{i,t} \tilde{k}_{ijt} = k_j$$

Notons $\tilde{N}(I \times T)$ le nuage des profils des lignes de $\tilde{k}_{(I \times T) \times J}$. Nous montrerons que le nuage $N'(I \times T)$ construit au paragraphe 5.2 s'en déduit par la translation k_j/k qui le centre à l'origine. Ceci est suffisant pour montrer que les facteurs sur $I \times T$ de l'analyse des correspondances du tableau $\tilde{k}$ se confondent avec les facteurs F'_s du nuage recentré. Pour obtenir les facteurs F_s , non recentrés, il suffira de mettre en éléments supplémentaires les lignes du tableau juxtaposé ainsi que celles du tableau somme.

Pour démontrer que $\tilde{N}(I \times T)$ se déduit de $N'(I \times T)$ par la translation k_j/k , il suffit de remarquer que ces deux nuages sont situés dans le même espace R_J (muni de la même métrique) et d'écrire les coordonnées d'un point (i,t) :

$$\text{dans } N'(I \times T) \quad (i,t) \rightsquigarrow k_{ijt}/k_{it} - k_{ij}/k_i$$

$$\text{dans } \tilde{N}(I \times T) \quad (i,t) \rightsquigarrow k_{ijt}/k_{it} - k_{ij}/k_i - k_j/k$$

Les poids affectés à ces points sont égaux, les deux nuages sont donc identiques.

5.3.2. deuxième technique

Dans cette seconde technique, on travaille directement sur le tableau juxtaposé et le tableau somme, mais, le programme diffère du programme d'analyse des correspondances. Dans ce dernier, les facteurs sur J sont calculés en diagonalisant une certaine matrice, les facteurs sur l'autre ensemble s'en déduisent par les formules de transition. Ici, la démarche est identique, mais les "facteurs sur J" sont calculés en diagonalisant la différence entre deux matrices : celle qui est diagonalisée dans l'analyse du tableau juxtaposé et celle qui est diagonalisée dans l'analyse du tableau somme.

La démonstration repose sur quatre points :

- . les projecteurs sur les axes d'inertie d'un nuage sont les vecteurs propres du produit de deux matrices : la matrice de sa forme d'inertie et la matrice de la métrique de l'espace ambiant (cf 2)
- . la métrique définie sur $\mathbb{R}_J$ est k/k_J pour le tableau juxtaposé et pour le tableau somme
- . d'après le théorème de Huyghens, l'inertie intra de N ($I \times T$) est la différence entre son inertie totale et son inertie inter (inertie de N(I) associée au tableau somme)
- . en analyse des correspondances, les facteurs définis sur les colonnes J se confondent, à la norme près, avec les projecteurs sur les axes d'inertie du nuage des lignes situé dans $\mathbb{R}_J$. (cf.2)

Les projecteurs sur les axes d'inertie intra de N ($I \times T$) s'obtiennent donc en diagonalisant la différence des deux matrices associées au tableau juxtaposé et au tableau somme. Ils se confondent, à la norme près, avec les facteurs sur J de l'analyse des correspondances du tableau k défini au paragraphe 5.3.1. Les facteurs F_s et F'_s s'en déduisent par les formules de projection.

Notons G_s le facteur sur J de norme $\sqrt{\lambda_s}$:

$$F_s(i) = (1/\sqrt{\lambda_s}) \sum_{j \in J} (k_{ij}/k_i) G_s(j)$$

$$F_s(i,t) = (1/\sqrt{\lambda_s}) \sum_{j \in J} (k_{ijt}/k_{it} - k_{ij}/k_i) G_s(j)$$

$$F'_s(i,t) = (1/\sqrt{\lambda_s}) \sum_{j \in J} (k_{ij}/k_i) G_s(j)$$

$$= F_s(i,t) - F_s(i)$$

6. LE NUAGE $N(J)$

Les facteurs sur J, G_s , de l'analyse du tableau $\tilde{k}_{(I \times T) \times J}$ sont, comme toujours en analyse des correspondances, les projections d'un nuage sur ces axes principaux d'inertie. Nous étudions ici ce nuage que nous notons $N(J)$. Il est situé dans l'espace $\mathbb{R}_{I \times T}$ muni de la métrique diagonale k/k_{IT} .

6.1. les coordonnées des points j

Les coordonnées du profil j d'une colonne du tableau $\tilde{k}$ s'écrivent, en prenant comme origine le centre de gravité (k_{IT}/k) de $N(J)$:

$$j \rightsquigarrow k_{IjT}/k_j - (k_{Ij}/k_j) (k_{IT}/k_I).$$

Ce sont les différences entre les coordonnées de deux points : le profil de la colonne j du tableau juxtaposé noté $j_{I \times T}$ et un point noté j_I de coordonnée $(k_{Ij}/k_j) (k_{IT}/k_I)$. Nous allons voir que ces points j_I représentent, dans l'espace $\mathbb{R}_{I \times T}$, le nuage des profils des colonnes du tableau somme.

6.2. nuage des profils du tableau somme

Calculons la distance entre les deux points j_I et j'_I associés aux colonnes j et j' et dont les coordonnées sont respectivement $(k_{Ij}/k_j) (k_{IT}/k_I)$ et $(k_{Ij'}/k_{j'}) (k_{IT}/k_I)$:

On a

$$\begin{aligned} D^2(j_I, j'_I) &= \sum_{i,t} \left(\frac{k_{ij}}{k_j} - \frac{k_{ij'}}{k_{j'}} \right)^2 \left(\frac{k_{it}}{k_i} \right)^2 \frac{k}{k_{it}} \\ &= \sum_i \left(\frac{k_{ij}}{k_j} - \frac{k_{ij'}}{k_{j'}} \right)^2 \frac{k}{k_i} \end{aligned}$$

$D(j_I, j'_I)$ est exactement la distance du χ^2 entre les profils des colonnes j et j' du tableau somme. L'ensemble des points $\{j_I, j \in J\}$ est donc isomorphe au nuage des profils des colonnes du tableau somme.

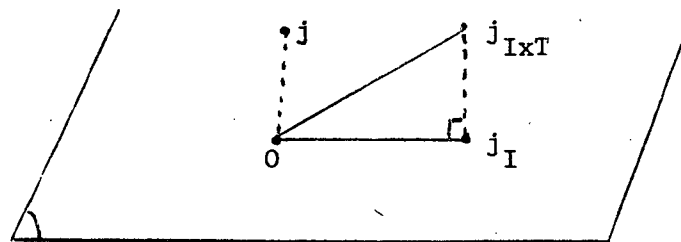
6.3. N(J) est un nuage recentré

Prenons comme origine O le centre de gravité du nuage $N(J)$;
 O est aussi le centre de gravité du nuage $N_{I \times T}(J)$ associé au tableau juxtaposé et du nuage $N_I(J)$ défini ci-dessus et représentant les colonnes du tableau somme dans $\mathcal{R}_{I \times T}$.

Entre les 3 points j , $j_{I \times T}$ et j_I , représentant la colonne j dans ces 3 nuages, on a la relation :

$$Oj = Oj_{I \times T} - Oj_I$$

$N(J)$ s'obtient donc en "recentrant" $N_{I \times T}(J)$ autour de $N_I(J)$. Montrons que $N(J)$ est la projection de $N_{I \times T}(J)$ sur l'orthogonal du sous espace engendré par $N_I(J)$:



sous-espace engendré par les Oj_I

Il suffit de montrer que pour tout j et tout j' , Oj est orthogonal à Oj'_I , i.e. que le produit scalaire $\langle Oj, Oj'_I \rangle$ est nul quels que soient j et j' .

$$\begin{aligned} \langle Oj, Oj'_I \rangle &= \sum_{i,t} \left(\frac{k_{ijt}}{k_j} - \frac{k_{ij} k_{it}}{k_i k_j} \right) \left(\frac{k_{ij'} k_{it}}{k_i k_{j'}} \right) \frac{k}{k_{it}} \\ &= \sum_i \left(\frac{k_{ij}}{k_j} \frac{k_{ij'}}{k_i} \frac{k}{k_j} - \frac{k_{ij} k_i k_{ij'} k}{k_i k_j k_i k_{j'}} \right) = 0 \end{aligned}$$

Le "recentrage" autour du nuage moyen associé au tableau somme a donc un aspect très différent pour les lignes et pour les colonnes. Notons l'analogie avec le centrage des variables en analyse en composantes principales qui se traduit par un changement d'origine pour les individus et par une projection sur l'orthogonal de la première bissectrice pour les variables.

6.4. interprétation des distances dans N(J)

Puisque le point j est la projection orthogonale du point $j_{I \times T}$ sur l'orthogonal de sous espace engendré par les O_{j_I} , on a la relation :

$$\|O_j\|^2 = \|O_{j_{I \times T}}\|^2 - \|O_{j_I}\|^2$$

De même le carré de la distance entre deux points j et j' dans $N(J)$ s'écrit en fonction des distances $D_{I \times T}$ et D_I induites respectivement par le tableau juxtaposé et le tableau somme :

$$D^2(j, j') = D_{I \times T}^2(j, j') - D_I^2(j, j')$$

Dans la distance (au carré) d'un point j au centre de gravité (ou entre deux points j et j') on soustrait de la distance (au carré) induite par le tableau juxtaposé, la distance (au carré) induite par le tableau somme. Un point j est donc éloigné du centre de gravité si l'écart de son profil à la marge est beaucoup plus grand dans le tableau juxtaposé que dans le tableau moyen. Deux points j et j' sont éloignés si l'écart entre les profils du tableau juxtaposé est beaucoup plus grand que l'écart entre les profils du tableau moyen.

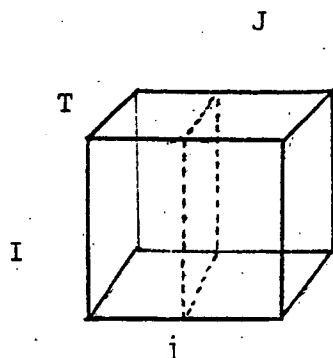
Remarquons que les projections du nuage $N_{I \times T}(J)$ associé au tableau juxtaposé sur les axes d'inertie de $N(J)$ se confondent avec les projections G_s de ce dernier. Les facteurs G_s sont donc aussi les projections de $N_{I \times T}(J)$ sur des axes particuliers mettant en évidence uniquement les écarts entre profils qui n'existent plus dans le tableau moyen.

6.5. interprétation en termes de correspondance ternaire

Le vecteur O_j peut aussi s'interpréter comme la différence entre deux mesures de probabilité sur $I \times T$ induites par $k_{I \times T}$. La première est la mesure de probabilité conditionnelle sachant j . La seconde est le produit d'une mesure sur I et d'une mesure sur T qui peuvent aussi s'écrire comme des probabilités conditionnelles :

$$j \rightsquigarrow P\{(i, t)/j\} - P\{i/j\} P\{t/i\}$$

L'analyse de $N(J)$ est donc l'analyse des écarts entre ces deux mesures induites sur $I \times T$. En termes de correspondance ternaire, la première est définie par une "tranche" du tableau ternaire $I \times J \times T$. Les deux autres sont définies par les marges sur $I \times J$ et $I \times T$ de ce tableau.



Correspondance $I \times J \times T$:

Le point j sera éloigné du centre de gravité si la "tranche" induite par j ne peut être reconstituée à partir des marges sur $I \times J$ et sur $I \times T$, s'il existe donc un terme d'interaction ternaire.

Remarquons que ce ne sont pas les termes d'interaction ternaire qui sont analysés. Dans ces derniers I, J, T jouent un rôle symétrique (cf.1), ce qui n'est pas le cas ici. Nous avons précisé que la comparaison s'orientait uniquement selon la dimension T et, ceci posé, nous avons encore dissymétrisé le problème sur I et sur J en faisant jouer un rôle différent aux lignes et aux colonnes.

7. ECART ENTRE LES MARGES SUR I DES TABLEAUX

Dans ce paragraphe, nous présentons la même méthode avec une approche totalement différente : comme une analyse de l'écart entre les données réelles et un tableau modèle construit sous certaines hypothèses. Ce point de vue apporte un autre éclairage aux résultats obtenus et ouvre la voie à une généralisation de l'analyse des correspondances à la comparaison de deux tableaux quelconques.

7.1. Le problème des écarts entre les marges

Généralement, les marges des tableaux que l'on compare ne sont pas proportionnelles, ce qui induit déjà une différence entre ces tableaux. Mais d'autres phénomènes plus complexes s'ajoutent à ce phénomène simple. Nous avons vu, au premier paragraphe, que dans certaines techniques de comparaison, on analysait essentiellement les écarts entre les marges des tableaux, pour peu que ces écarts soient relativement importants. D'où la question : comment analyser les différences entre les tableaux en évitant que l'écart entre leur marge ne masque des phénomènes plus fins ?

La solution proposée ici est d'éliminer totalement l'effet des écarts entre les marges sur I : en construisant un modèle où les tableaux ne diffèrent que par leur marge sur I, et en analysant la différence entre les données réelles et ce modèle.

7.2. Le modèle

Si les tableaux ne diffèrent que par leur marge sur I, chaque tableau t se déduirait du tableau moyen en multipliant tous les termes de la ligne i par le rapport k_{it}/k_i des marges sur I. On aurait :

$$k_{ijt} = k_{ij} (k_{it}/k_i) \text{ pour tout } i, j, t$$

Les profils des lignes homologues seraient alors identiques dans tous les tableaux :

$$k_{ijt}/k_{it} = k_{ij}/k_i$$

Le tableau modèle correspondant à cette hypothèse, défini sur le produit $(I \times T) \times J$, a pour terme général :

$$k'_{ijt} = k_{ij} (k_{it}/k_i)$$

Ce tableau a les mêmes marges que le tableau juxtaposé :

$$k'_{it} = \sum_j k'_{ijt} = k_{it}$$

$$k'_j = \sum_{i,t} k'_{ijt} = k_j$$

7.3. Analyse de la différence entre données et modèle

Nous voulons maintenant analyser l'écart entre le tableau juxtaposé k_{ijt} et le modèle k'_{ijt} construit sous l'hypothèse que les tableaux t ne diffèrent que par leur marge sur I .

Le modèle et le tableau réel ont les mêmes marges : les métriques induites sur $R_{I \times T}$ et R_J sont identiques et les profils des lignes et des colonnes sont construits de la même façon.

On analysera donc :

- Dans $R_{I \times T}$ muni de la métrique diagonale k/k_{IT} , le nuage des différences entre les profils k_{ijt}/k_j et $k'_{ijt}/k_j = k_{ij}/k_i$. C'est exactement le nuage $N'(I \times T)$ introduit au paragraphe 6.

- Dans R_J muni de la métrique diagonale k/k_J , le nuage des différences entre les profils k_{ijt}/k_j et $k'_{ijt}/k_j = (k_{ij}/k_j)(k_{it}/k_i)$. C'est le nuage $N(J)$ introduit au paragraphe 7.

L'analyse de l'écart des données au modèle indiqué conduit donc exactement aux mêmes résultats que la maximisation de l'inertie intra.

Remarque : L'analyse des correspondances d'un tableau de contingence P_{ij} est l'analyse de l'écart entre la loi réelle P_{ij} et la loi produit $P_i P_j$, qui constitue un modèle construit sous l'hypothèse d'indépendance. La méthode proposée apparaît ainsi comme une généralisation de l'analyse des correspondances à un modèle différent de celui donné par le produit des lois marginales.

8. REPRESENTATION SIMULTANEE ET INTERPRETATION DES RESULTATS

Avant d'indiquer les formules de transition de la représentation simultanée des ensembles $I \times T$ et J , rappelons la signification des facteurs définis sur chacun de ces ensembles.

8.1. Facteurs sur $I \times T$: F_s et F'_s

Pour définir ces facteurs, nous avons introduit deux nuages $N(I \times T)$ et $N'(I \times T)$. Le premier est le nuage des profils des lignes du tableau juxtaposé, le second se déduit du premier en recentrant à l'origine chaque sous nuage des lignes homologues $\{(i,t), t \in T\}$.

Les facteurs F_s et F'_s sont les projections de ces deux nuages sur des axes, qui sont à la fois les axes d'inertie "intra" de $N(I \times T)$ et les axes principaux d'inertie de $N'(I \times T)$.

Entre F_s et F'_s , on a les relations :

$$F'_s(i,t) = F_s(i,t) - F_s(i)$$

où $F_s(i)$ est la projection du profil de la ligne i du tableau somme, situé au centre de gravité des (i,t) .

Ces facteurs sont centrés :

$$\sum_{i,t} k_{it} F'_s(i,t) = \sum_{i,t} k_{it} F_s(i,t) = 0$$

et de plus

$$\sum_i k_i F_s(i) = 0 \text{ et } \sum_t k_{it} F'_s(i,t) = 0$$

8.2. Contribution absolue et qualité de représentation des (i,t)

Les contributions absolues et les qualités de représentation des éléments sont des aides à l'interprétation qui nous paraissent maintenant indispensables en analyse des correspondances. Ces deux notions s'appliquent ici sans aucune difficulté puisque nous projetons des nuages sur des axes orthogonaux deux à deux.

La contribution absolue d'un élément (i,t) à l'inertie λ_s du facteur d'ordre s s'écrit en fonction de F'_s et de F_s . Son expression est plus simple avec le premier qu'avec le second où il faut écrire l'inertie "intra" de (i,t) .

$$\begin{aligned} \text{CTR}(i,t) &= (k_{it}/k) \{ F'_s(i,t) \}^2 \\ &= (k_{it}/k) \{ F_s(i,t) - F_s(i) \}^2 \end{aligned}$$

Un élément i du tableau somme ne contribue pas à l'inertie du facteur, mais on appellera contribution de i la contribution du nuage $N(i \times T)$:

$$\text{CTR}(i) = \sum_{t \in T} \text{CTR}(i,t)$$

De même, on appellera contribution du tableau t à l'inertie du facteur, la contribution de ses éléments :

$$\text{CTR}(t) = \sum_{i \in I} \text{CTR}(i,t)$$

Un élément (i,t) contribuera beaucoup à l'inertie si le profil de cette ligne est très différent du profil de la ligne i du tableau moyen, et si son poids k_{it} est assez grand. La contribution d'un élément i à l'inertie sera grande si les profils des lignes $\{(i,t), t \in T\}$ diffèrent entre eux. Celle d'un tableau t sera grande si les profils des lignes de ces tableaux diffèrent de celles du tableau moyen.

Les qualités de représentation des (i,t) peuvent être calculées dans le nuage initial $N(I \times T)$ ou dans le nuage recentré $N'(I \times T)$. Dans le premier cas, c'est la qualité de représentation de l'écart de (i,t) à la moyenne générale, elle n'atteint pas forcément 1 pour l'ensemble des facteurs. Dans le second, c'est celle de l'écart de (i,t) au centre de gravité i de $N(i \times T)$; elle atteint 1 pour l'ensemble des facteurs.

8.3. Transition de G_s vers F'_s et F_s

En projetant le nuage $N(I \times T)$, on a :

$$F_s(i,t) = (1/\sqrt{\lambda_s}) \sum_j (k_{ijt}/k_{it}) G_s(j)$$

où λ_s est à la fois l'inertie de G_s et de F'_s et l'inertie intra de F_s .

Cette formule est la formule de transition du tableau juxtaposé : le point $F_s(i,t)$ qui est, à $1/\sqrt{\lambda_s}$ près, au centre de gravité des $G_s(j)$ est situé du même côté de l'axe que les points j auxquels il s'associe le plus.

De même, la projection $F_s(i)$ du profil de la ligne i du tableau somme s'écrit :

$$F_s(i) = (1/\sqrt{\lambda_s}) \sum_j (k_{ij}/k_i) G_s(j)$$

Le facteur F'_s , projection du nuage recentré se déduit de G_s par :

$$F'_s(i,t) = (1/\sqrt{\lambda_s}) \sum_j (k_{ijt}/k_{it} - k_{ij}/k_i) G_s(j)$$

On trouvera donc dans la représentation simultanée de $I \times T$ et de J par les facteurs F'_s et G_s aux extrémités des axes :

- . les lignes (i,t) qui s'éloignent le plus de la ligne moyenne i ,
- . les colonnes j qui participent le plus à cet écart à la moyenne et qui expliquent la position des lignes : $F'_s(i,t)$ est situé du même côté de l'axe que les j auxquels il s'associe plus que la moyenne i des lignes homologues, et du côté opposé des j auxquels il s'associe moins que i .

8.4. Transition de F_s et F'_s vers G_s

La formule de transition de F'_s vers G_s peut s'écrire en utilisant, par exemple, le fait que F'_s et G_s sont les facteurs de l'analyse des correspondances des tableaux $\tilde{k}_{ijt}$ (cf. 5.3.1.) :

$$G_s(j) = (1/\sqrt{\lambda_s}) \sum_{i,t} \{ (k_{ijt}/k_j) - (k_{ij}/k_i) / (k_{it}/k_j) \} F'_s(i,t)$$

Cette formule se simplifie et devient la formule de transition du tableau juxtaposé. En effet, les projections de $N(J)$ sur ses axes d'inertie se confondent avec les projections du nuage associé au tableau juxtaposé (cf. § 6.3.). D'où :

$$G_s(j) = (1/\sqrt{\lambda_s}) \sum_{i,t} (k_{ijt}/k_j) F'_s(i,t)$$

Remarquons que la formule de transition duale est vérifiée pour F_S , mais non pour F'_S .

Il est facile de vérifier que la première formule de transition reste valable pour F_S en remplaçant $F'_S(i,t)$ par $F_S(i,t) + F_S(i)$. Mais la seconde ne l'est plus, on a donc :

$$G_S(j) = (1/\sqrt{\lambda_S}) \sum_{i,t} \{ (k_{ijt}/k_j) - (k_{ij}/k_i)(k_{it}/k_j) \} F_S(i,t)$$

Aux extrémités des axes G_S sont situés les éléments j qui sont loin du centre de gravité du nuage ; i.e. les j tels que k_{ijt} ne se déduit pas des marges k_{it} et k_{ij} ; tels que la colonne du tableau juxtaposé diffère beaucoup de la colonne du modèle. Les facteurs $F_S(i,t)$ expliquent bien les positions des $G_S(j)$: $G_S(j)$ est situé sur l'axe du côté des (i,t) auxquels j s'associe plus dans le tableau réel que dans le modèle, et du côté opposé des (i,t) auxquels j s'associe moins que dans le modèle.

8.5. Éléments supplémentaires

Les formules de transition permettent de projeter des lignes ou des colonnes supplémentaires. En ligne, on projettera les marges sur J de chaque tableau, ou encore un tableau tout entier.

9. LES DEUX ANALYSES DES LIGNES ET DES COLONNES

Nous avons déjà souligné plusieurs fois que lignes et colonnes ne jouaient pas du tout le même rôle dans l'analyse que nous proposons.

Il est clair qu'une seconde analyse est possible en inversant le rôle des lignes et des colonnes. Les résultats s'interprètent de la même façon, mais ces deux analyses sont différentes et se complètent. Le cas extrême où toutes les lignes homologues sont proportionnelles, le montre bien. L'inertie intra du nuage $N(IxT)$ est alors nulle et dans la première analyse tous les facteurs sont nuls. Si les lignes sont proportionnelles,

les colonnes ne peuvent l'être, à moins d'égalité des tableaux, et dans la seconde analyse les facteurs ne peuvent être tous nuls.

Si, quelquefois, une seule des deux analyses présente de l'intérêt, la plupart du temps les deux analyses seront nécessaires pour avoir une vue assez complète des différences entre les tableaux.

On peut se demander s'il serait possible de faire jouer un rôle symétrique aux lignes et aux colonnes, ce qui serait intéressant, surtout pour éliminer les écarts entre les marges. Il semble bien que cette approche ne le permette pas.

Dans le cas de tableau ternaire, où l'une des dimensions ne s'impose pas pour la comparaison, on peut aussi permuter T avec I et J, ce qui permet 6 analyses différentes...

10. EXEMPLE

10.1. Les données (cf. tableau 1)

Ce sont deux petits tableaux de contingence répartissant la population française sortie du système éducatif en 1973 et ayant trouvé un emploi la même année suivant les deux critères : niveau de diplôme et catégorie d'emploi. L'un concerne la population masculine, l'autre la population féminine. Ces tableaux sont extraits de [5] et utilisés dans [6] pour illustrer l'intérêt de quelques indices de comparaison concernant les marges et les facteurs des tableaux.

Nous ferons deux analyses différentes avec la méthode proposée :

- 1) en mettant en ligne les niveaux de diplôme
- 2) en mettant en ligne les emplois occupés.

10.2. Comparaison des profils d'emploi à diplôme égal

Dans cette analyse, nous mettons en évidence pour chaque emploi, la différence entre les diplômes des deux populations. Partant d'un autre point de vue, nous analysons l'écart du tableau croisant diplômes-sexe

Emplois occupés*	Niveaux de diplôme								Total
	SD sans diplôme	BEP BEPC	CAP BEP/CAP	BAC BAC général	BTE BAC technique	DEG DEUG/ENT	DUT DUT/BTS/ Santé	SUP SUP	
1. Agriculteurs (AGR)	15 068	2 701	5 709	297	1 242	-	322	-	25 339
2. Ingénieurs (ING)	-	337	309	917	-	308	-	4 383	6 254
3. Techniciens (TEC)	302	1 697	2 242	1 969	1 399	357	1 943	381	10 290
4. Ouvriers qualifiés (OQ)	10 143	3 702	30 926	314	1 861	-	-	337	47 283
5. Ouvriers non qualifiés (ONQ)	59 394	8 087	17 862	2 887	1 696	-	-	323	90 249
6. Cadres supérieurs (CSUP)	596	298	892	1 227	298	2 362	318	6 781	12 772
7. Cadres moyens (CM)	2 142	2 801	672	6 495	924	2 807	2 301	4 030	22 172
8. Employés qualifiés (EQ)	5 445	7 348	4 719	4 353	1 280	614	982	-	24 741
9. Employés non qualifiés (ENQ) ...	4 879	4 987	1 514	3 478	886	1 326	-	661	17 431
Total	97 969	31 958	64 845	21 937	9 586	7 774	5 866	16 896	

* Source : "Bilan formation-emploi 1973", CEREQ, INSEE, SEIS, volume D 59 des Collections de l'INSEE, p. 102 et 103.

ELEVES SCOLARISES EN 1972-1973, SORTIS DU SYSTEME EDUCATIF EN 1973 ET AVANT TROUVE UN EMPLOI - SEXE FEMININ*

Emplois occupés	Niveaux de diplôme								Total
	SD sans diplôme	BEP BEPC	CAP BEP/CAP	BAC BAC général	BTE BAC technique	DEG DEUG/ENT	DUT DUT/BTS Santé	SUP SUP	
1. Agriculteurs (AGR)	5 089	1 212	1 166	-	-	-	-	-	7 467
2. Ingénieurs (ING)	-	-	-	316	-	-	304	1 033	1 653
3. Techniciens (TEC)	281	-	320	320	283	-	683	-	1 887
4. Ouvriers qualifiés (OQ)	7 470	1 859	4 017	1 752	657	-	285	-	16 040
5. Ouvriers non qualifiés (ONQ)	29 997	4 334	4 538	1 882	-	-	-	-	40 751
6. Cadres supérieurs (CSUP)	-	-	-	2 236	595	911	569	6 788	11 099
7. Cadres moyens (CM)	1 577	1 806	4 549	17 063	875	4 152	15 731	3 991	49 744
8. Emplois qualifiés (EQ)	21 616	19 915	32 452	16 137	5 865	1 256	3 332	1 286	101 859
9. Emplois non qualifiés (ENQ)	19 849	7 325	6 484	5 111	898	294	635	-	40 596
Total	85 879	36 451	53 526	44 817	9 173	6 613	21 539	13 098	

Tableau 1

avec emploi au tableau modèle de même taille où les deux tableaux concernant les hommes et les femmes ne diffèrent que par la répartition marginale des diplômés.

10.2.1. Inertie

L'inertie totale de cette analyse qui est donc l'inertie "intra" du nuage des profils d'emploi des diplômés-sexe vaut 0.38. Cette inertie est importante par rapport à l'inertie totale de ce nuage, 1.32, donnée par l'analyse des correspondances du tableau juxtaposé.

La répartition des emplois, à diplôme égal, diffère donc chez les hommes et chez les femmes.

10.2.2. Premier facteur (cf. tableau 2)

Le premier facteur extrait 60 % de l'inertie. Le diplôme qui contribue le plus à l'inertie de ce facteur est le CAP dont la contribution à l'inertie est de 29 % pour le CAP-homme et 35 % pour le CAP-femme, donc 64 % en tout. C'est le diplôme pour lequel la répartition des emplois est la plus différente chez les hommes et chez les femmes. Cette différence est entièrement expliquée par le premier facteur ($COR=0.97$) qui explique bien aussi la différence entre les emplois des BEP et des BAC Techniques, qui sont aussi des diplômes de niveau moyen. Remarquons que dans le facteur recentré ($F'1$) traduisant les écarts des profils d'emploi des diplômés - sexe au profil d'emploi moyen du même diplôme, les diplômés - hommes sont tous situés à droite et les diplômés femmes à gauche. Le premier facteur traduit donc une tendance générale de l'ensemble des diplômés.

Les emplois les plus opposés sur ce premier facteur sont : d'une part, les "employés qualifiés" situés à gauche et donc occupés proportionnellement plus par les femmes ; et d'autre part, les emplois d'ouvriers qualifiés et de techniciens qui sont préférés par les hommes.

10.2.3. Deuxième facteur (cf. tableau 2)

Ce second facteur ne représente plus que 15 % de l'inertie. Il montre une opposition entre les emplois occupés par les hommes et par les femmes chez les possesseurs du BAC ou d'un DUT. Pour ces emplois, les hommes

Facteurs recentrés sur IxT = Diplôme-Sexe

H Homme

F Femme

I	I1	GLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR	4#F	COR	CTR
1	Hsd	991	186	91	329	569	86	-143	108	64	224	265	232	-96	49	50
2	Hbep	872	61	32	412	822	44	66	21	4	68	23	7	-34	6	2
3	Hcap	997	123	190	742	923	288	-40	3	3	-187	59	108	86	12	26
4	Hbac	930	42	29	257	246	12	409	622	117	83	25	7	-99	37	12
5	Hbte	851	18	25	626	748	30	211	85	14	96	18	4	12	0	0
6	Hdes	407	15	12	107	36	1	332	342	27	-52	9	1	81	20	3
7	Hdut	961	11	86	442	66	9	1427	685	380	629	133	109	479	77	73
8	Hsup	989	32	41	117	28	2	289	169	45	-225	103	40	-583	689	312
9	Fsd	991	163	104	-374	569	98	164	108	73	-255	265	265	110	49	57
10	Fbep	873	69	28	-360	823	38	-57	21	4	-59	23	6	31	6	2
11	Fcap	997	101	230	-898	923	349	50	3	4	228	59	130	-103	12	32
12	Fbac	931	85	14	-125	246	6	-199	622	57	-39	25	3	49	38	6
13	Fbte	851	17	26	-653	748	32	-219	85	14	-99	18	4	-12	0	0
14	Fdes	409	13	14	-125	36	1	-390	344	32	62	9	1	-94	20	3
15	Fdut	960	41	23	-119	66	3	-388	684	103	-170	133	30	-130	77	20
16	Fsup	989	25	53	-149	28	2	-372	169	58	291	103	52	753	689	403
		1527928.0	1000			1000			1000			1000				1000

Facteurs sur J = emplois

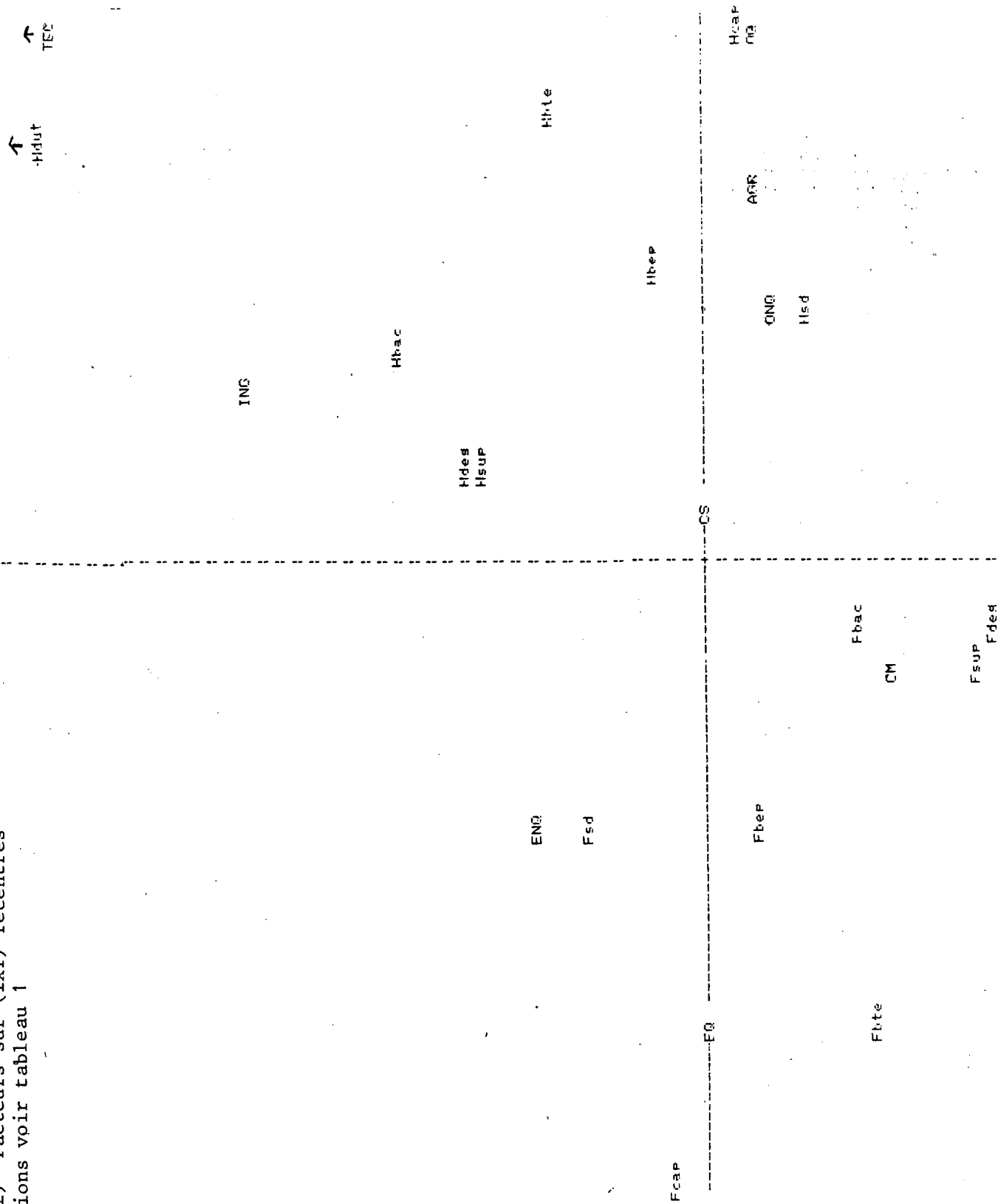
I	J1	GLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR	4#F	COR	CTR
1	AGR	982	62	54	497	744	66	-68	14	5	263	208	106	-72	16	9
2	ING	986	15	88	240	25	4	615	168	95	-458	93	78	-1256	700	677
3	TEC	983	23	140	692	204	47	1261	677	615	417	74	99	255	28	43
4	OO	986	120	201	695	749	247	-74	9	11	-349	190	364	156	38	84
5	ONQ	989	248	92	328	751	114	-80	45	27	147	151	133	-76	42	42
6	CS	504	45	24	68	23	1	-3	0	0	93	43	10	297	438	114
7	CM	736	136	40	-133	159	10	-248	548	142	2	0	0	-56	29	13
8	EQ	979	240	277	-657	973	442	12	0	1	47	5	13	20	1	3
9	ENQ	934	110	85	-382	492	69	237	188	104	-266	239	195	68	15	14
		1527928.0	1000			1000			1000			1000				1000

Comparaison des pourcentages d'emplois à diplôme fixé

I = diplôme J = emploi T = sexe

TABLEAU 2

Comparaison des pourcentages d'emploi à diplôme
fixé
Plan (1,2) Facteurs sur (Ixt) recentrés
Abréviations voir tableau 1



sont plutôt technicien ou ingénieur, et les femmes plutôt cadre moyen.

10.3. Comparaison des profils de diplôme à emploi égal

Dans cette seconde analyse, nous mettons en évidence, pour chaque emploi, les différences entre les niveaux de diplômes des hommes et des femmes. Dans la comparaison des deux tableaux, nous éliminons l'effet de l'écart entre les répartitions marginales des emplois. Cet écart est beaucoup plus important que celui des répartitions marginales des diplômes.

10.3.1. Inertie

L'inertie totale de cette analyse est quatre fois plus faible que celle de la précédente : la répartition des diplômes à emploi égal diffère beaucoup moins entre les deux sexes que la répartition des emplois à diplôme égal.

10.3.2. Le premier facteur

Il extrait 46 % de l'inertie. Sur ce facteur, les emplois pour lesquels la répartition des diplômes s'oppose le plus dans les deux sexes sont les emplois d'ingénieur, technicien et cadre moyen. Les 3 emplois-femmes sont tous situés à gauche des emplois-hommes identiques. Les femmes occupant ces types d'emplois ont beaucoup plus souvent que les hommes un DUT ; l'opposition du DUT aux autres niveaux de diplôme prédomine sur le facteur associé.

10.3.3. Le deuxième facteur

Il représente 31 % de l'inertie et oppose les profils de diplôme chez les ouvriers qualifiés : les hommes ont le plus souvent un CAP ; chez les femmes on trouve en plus grande proportion que chez les hommes des DUT, des BAC, ou des sans diplômes.

Facteurs recentrés sur IxT

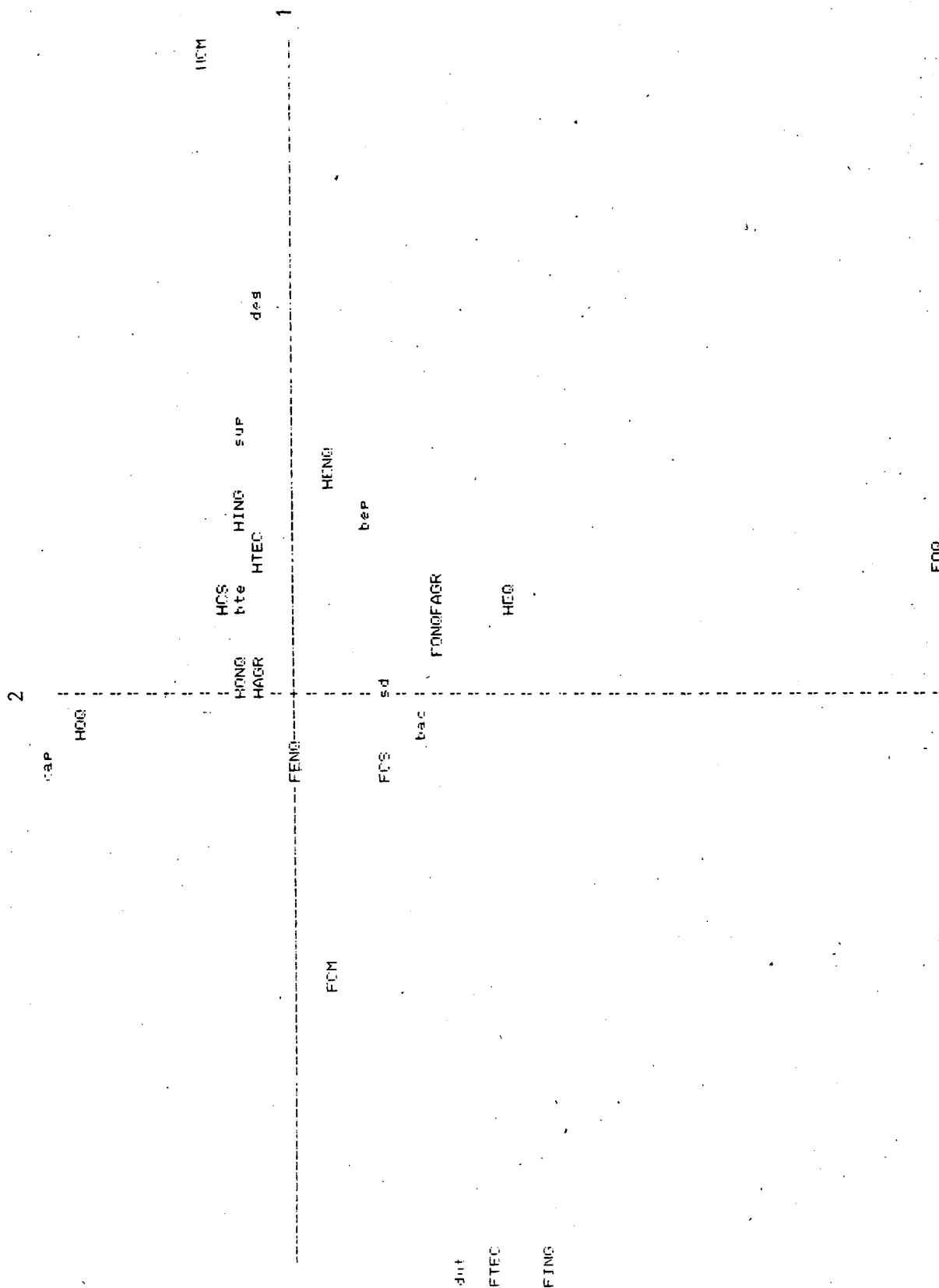
	I1	Q1T	P01D	I1NR	1#F	C0R	C1TR	2#F	C0R	C1TR	3#F	C0R	C1TR	4#F	C0R	C1TR
1	HAGR	604	48	4	-24	89	1	44	265	3	-15	36	1	39	214	10
2	HING	988	12	5	181	842	9	71	130	2	21	11	0	-12	5	0
3	HTEC	831	19	5	127	728	7	42	78	1	-23	25	1	1	0	0
4	HOO	987	90	62	-41	27	4	248	952	185	-21	8	4	0	0	0
5	HONQ	937	171	11	-14	34	1	70	784	28	-15	41	4	22	78	12
6	HCS	974	24	37	81	45	4	89	55	6	-124	109	34	-330	765	375
7	HCM	999	42	266	755	956	554	99	16	14	126	27	60	16	0	1
8	HEQ	852	47	53	105	103	12	-240	543	91	-142	191	86	-39	15	11
9	HENQ	989	34	85	246	254	47	-29	4	1	-400	673	482	117	58	65
10	FAGR	604	14	12	86	89	2	-147	266	10	55	36	4	-132	213	35
11	FING	987	3	18	-682	841	34	-268	130	8	-77	11	2	51	5	1
12	FTEC	833	4	25	-689	728	39	-226	79	6	129	26	5	-7	0	0
13	FOQ	987	30	181	123	27	11	-731	952	545	66	8	12	4	0	0
14	FOHQ	938	77	25	32	34	2	-154	784	62	36	42	9	-48	78	26
15	FCS	974	21	42	-91	45	4	-101	55	7	144	109	39	381	765	431
16	FCM	999	94	119	-336	956	247	-43	16	6	-55	27	27	-6	0	1
17	FEQ	852	193	13	-25	103	3	59	543	22	35	191	21	10	15	3
18	FENQ	989	77	37	-106	254	21	13	4	0	175	673	210	-50	58	29
I		1527930.0	1000			1000			1000			1000				1000

Facteurs sur J

I	J1	Q1	P1	O1	I1	N1	R1	1#F	C1	O1	R1	C1	T1	R1	2#F	C1	O1	R1	C1	T1	R1	3#F	C1	O1	R1	C1	T1	R1	4#F	C1	O1	R1	C1	T1	R1		
1	sd	995	348	94				19	14	3					-104	430	128					105	432	344					-54	119	150						
2	bep	866	130	85				186	558	103					-79	105	28					-111	202	145					-7	1	1						
3	cap	999	224	240				-98	98	51					301	898	681					9	1	2					-12	2	6						
4	bac	956	126	49				-28	23	2					-143	572	88					-58	97	40					98	264	172						
5	bte	683	36	28				104	149	9					58	46	4					-51	38	9					181	450	165						
6	des	954	27	115				420	443	111					53	8	3					-373	353	342					-243	150	229						
7	dut	998	52	309				-708	897	602					-185	61	60					-148	40	103					11	0	1						
8	sup	981	57	81				298	667	117					66	33	8					57	24	16					185	257	275						
I		1527930.0	1000							1000							1000							1000											1000		

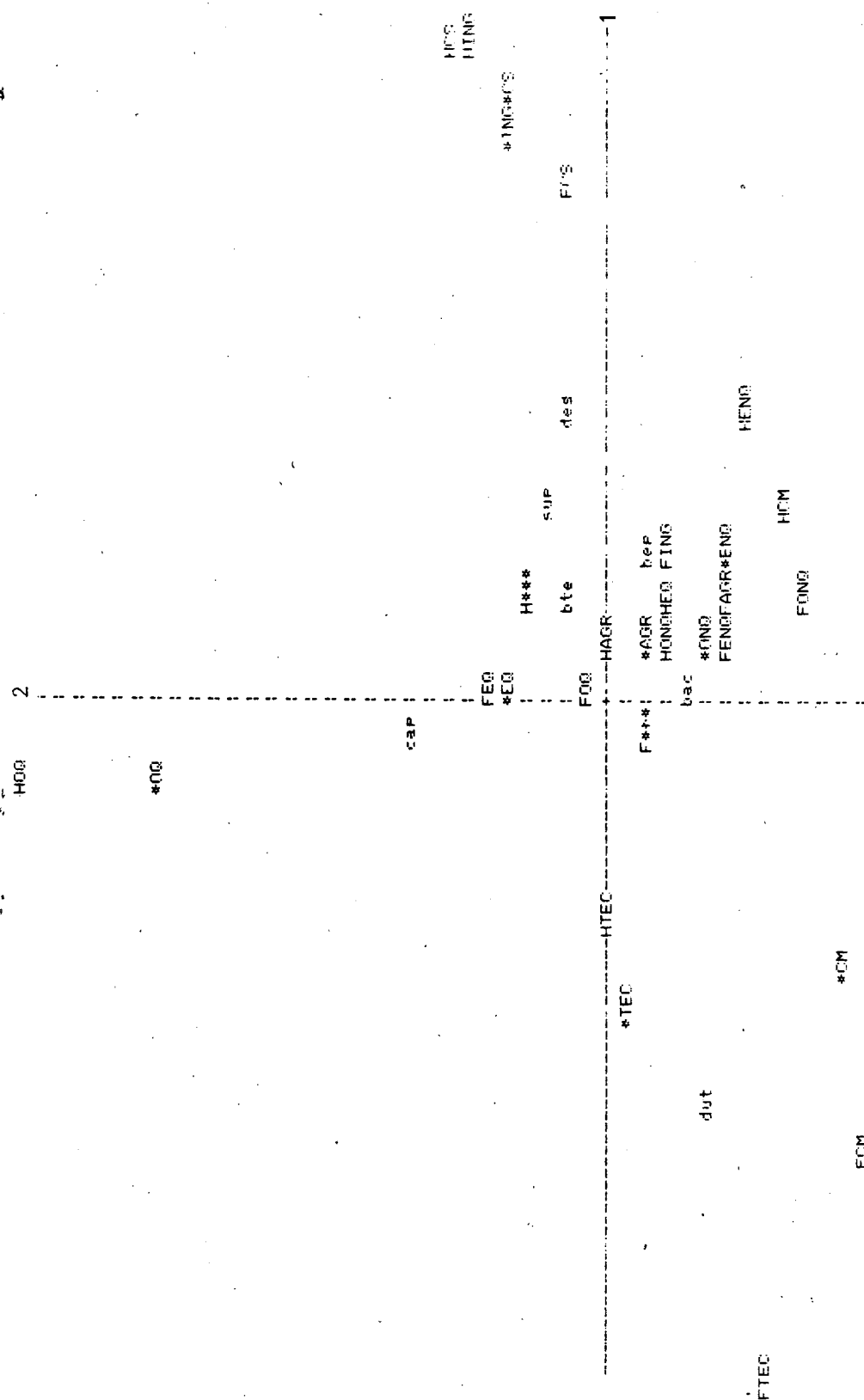
Comparaison des pourcentages de diplômes à emploi fixé
 I = emploi J = diplômes T = sexe

TABLEAU 3



Comparaison des pourcentages de diplômes à emploi fixé. Plan (1,2). Les facteurs sur Ixt = Emploi X Sexe sont les facteurs recentrés

GRAPHIQUE 3 Pour les symboles, voir tableau 1



Comparaison des profils des diplômés à emploi fixé

Plan (1,2) Facteurs sur IxT non recentrés

Abréviations : voir tableau 1.

Les lignes du tableau somme sont précédées d'une étoile

Les centres de gravité des 2 tableaux sont notés H*** et F***

11. CONCLUSION

Cette méthode répond à un besoin souvent exprimé par des utilisateurs : celui d'une technique permettant d'analyser systématiquement les différences entre plusieurs tableaux de fréquence.

On a vu, sur le petit exemple, que les écarts entre profils des lignes homologues, même très faibles, sont bien mis en évidence, leur signification est claire puisque leur distance est la distance du χ^2 . La représentation simultanée des colonnes permet d'interpréter complètement les résultats.

La non symétrie des deux analyses qui permutent lignes et colonnes est très visible dans cet exemple.

Aucune des techniques indiquées au troisième paragraphe ne permettait d'aboutir à ce résultat : la juxtaposition met en évidence essentiellement les dimensions communes des tableaux ; les autres techniques ne permettent pas de comparer plus de deux tableaux ; les distances ne s'interprètent pas en termes de distance du χ^2 ; dans la représentation simultanée, les facteurs ne vérifient pas de formules barycentriques exactes. Un vide est donc comblé. Remarquons que la complexité des calculs est celle d'une analyse des correspondances, donc très faible.

Pour comparer de manière cohérente et très complète plusieurs tableaux, on pourra faire : l'analyse des tableaux juxtaposés en lignes ou en colonnes avec la somme en principal ou en supplémentaire ; le calcul des indices globaux de comparaison des marges et des facteurs (cf. 6) ; les deux analyses intra.

BIBLIOGRAPHIE

- [1] BENER A.
Décomposition des interactions dans une correspondance multiple
Cahiers de l'analyse des données - 1982 - n°1.
- [2] BENZECRI J.P. et BENZECRI F.
Pratique de l'analyse des données
Dunod - 1980.
- [3] CAZES P.
Note sur les éléments supplémentaires en analyse de correspon-
dances
II Tableaux multiples
Cahiers de l'analyse des données n°2 - 1982.
- [4] CLAPIER P. et TABARD N.
Transformation de la morphologie sociale des communes et varia-
tion des consommations
Consommation n°2 - 1981.
- [5] DOMENGES D. et VOLLE M.
Analyse factorielle sphérique : une exploration
Annales de l'INSEE n°35 - 1979.
- [6] ESCOPIER B.
Quelques indices pour comparer des tableaux de contingence
Statistique et analyse des données 1981 n°1.
- [7] YAGOLNITZER E.
Comparaison de deux correspondances entre les mêmes ensembles
Cahier de l'analyse des données n°3 - 1977.
- [8] GOUDARD J. et GRELET Y.
Les lycéens du second cycle: Comparaison entre filles et garçons
Cahier de l'analyse des données n°3 - 1977.

Liste des Publications Internes IRISA

- PI 150 **Construction automatique et évaluation d'un graphe d'«implication» issu de données binaires, dans le cadre de la didactique des mathématiques**
H. Rostam , 112 pages ; Juin 1981
- PI 151 **Réalisation d'un outil d'évaluation de mécanismes de détection de pannes]-Projet Pilote SURF**
B. Decouty, G. Michel, C. Wagner, Y. Crouzet , 59 pages ; Juillet 1981
- PI 152 **Règle maximale**
J. Pellaumail , 18 pages ; Septembre 1981
- épuisée PI 153 **Corrélation partielle dans le cas « qualitatif »**
I.C. Lerman , 125 pages ; Octobre 1981
- épuisée PI 154 **Stability analysis of adaptively controlled, not-necessarily minimum phase systems with disturbances**
Cl. Samson , 40 pages ; Octobre 1981
- PI 155 **Analyses d'opinions d'instituteurs à l'égard de l'appropriation des nombres naturels par les élèves de cycle préparatoire**
R. Gras , 37 pages ; Octobre 1981
- PI 156 **Récursion induction principe revisited**
G. Boudol, L. Kott , 49 pages ; Décembre 1981
- PI 157 **Loi d'une variable aléatoire à valeur R^+ réalisant le minimum des moments d'ordre supérieur à deux lorsque les deux premiers sont fixés**
M. Kowalowka, R. Marie , 8 pages ; Décembre 1981
- épuisée PI 158 **Réalisations stochastiques de signaux non stationnaires, et identification sur un seul échantillon**
A. Benveniste J.J. Fuchs , 33 pages ; Mars 1982
- PI 159 **Méthode d'interprétation d'une classification hiérarchique d'attributs-modalités pour l'«explication» d'une variable ; application à la recherche de seuil critique de la tension artérielle systolique et des indicateurs de risque cardiovasculaire**
B. Tallur , 34 pages ; Janvier 1982
- PI 160 **Probabilité stationnaire d'un réseau de files d'attente multiclasse à serveur central et à routages dépendant de l'état**
L.M. Le Ny , 18 pages ; Janvier 1982
- épuisée PI 161 **Détection séquentielle de changements brusques des caractéristiques spectrales d'un signal numérique**
M. Basseville, A. Benveniste , pages ; Mars 1982
- PI 162 **Actes regroupés des journées de Classification de Toulouse (Mai 1980), et de Nancy (Juin 1981)**
I.C. Lerman , 304 pages ;
- PI 163 **Modélisation et Identification des caractéristiques d'une structure vibratoire : un problème de réalisation stochastique d'un grand système non stationnaire**
M. Prévosto, A. Benveniste, B. Barnouin , 46 pages ; Mars 1982
- PI 164 **An enlarged definition and complete axiomatization of observational congruence of finite processes**
Ph. Darondeau , 45 pages ; Avril 1982
- épuisée PI 165 **Accès vidéotex à une banque de données médicales**
A. Chauffaut, M. Dragone, R. Rivoire, J.M. Roger , 25 pages ; Mai 1982
- PI 166 **Comparaison de groupes de variables définies sur le même ensemble d'individus**
B. Escofier, J. Pages , 115 pages ; Mai 1982
- PI 167 **Transport en circuits virtuels internes sur réseau local et connexion Transpac**
M. Tournois, R. Trépos , 90 pages ; Mai 1982
- PI 168 **Impact de l'intégration sur le traitement automatique de la parole**
P. Quinton , 14 pages ; Mai 1982
- PI 169 **A systolic algorithm for connected word recognition**
J.P. Banâtre, P. Frison, P. Quinton , 13 pages ; Mai 1982
- PI 170 **A network for the detection of words in continuous speech**
J.P. Banâtre, P. Frison, P. Quinton , 24 pages ; Mai 1982
- PI 171 **Le langage ADA : Etude bibliographique**
J. André, Y. Jégou, M. Raynal , 12 pages ; Juin 1982
- épuisée PI 172 **Comparaison de groupes de variables : 2ème partie : un exemple d'application**
B. Escofier, J. Pajès , 37 pages ; Juillet 1982
- PI 173 **Unfold-fold program transformations**
L. Kott , 29 pages ; Juillet 1982
- PI 174 **Remarques sur les langages de parenthèses**
J.M. Autebert, J. Beauquier, L. Boasson, G. Senizergues , 20 pages ; Juillet 1982
- PI 175 **Langages de parenthèses, langages N.T.S. et homomorphismes inverses**
J.M. Autebert, L. Boasson, G. Senizergues , 26 pages ; Juillet 1982
- PI 176 **Tris pour machines synchrones ou Baudet Stevenson revisited**
R. Rannou , 26 pages ; Juillet 1982
- PI 177 **Un nouvel algorithme de classification hiérarchique des éléments constitutifs de tableau de contingence basé sur la corrélation**
B. Tallur , Juillet 1982 ;
- PI 178 **Programmes d'analyse des résultats d'une classification automatique**
I.C. Lerman et collaborateurs , 79 pages ; Septembre 1982
- PI 179 **Attitude à l'égard des mathématiques des élèves de sixième**
J. Degouys, R. Gras, M. Postic , 29 pages ; Septembre 1982
- épuisée PI 180 **Traitements de textes et manipulations de documents : bibliographie analytique**
J. André , 20 pages ; Septembre 1982

- PI 181 **Algorithme assurant l'insertion dynamique d'un processeur autour d'un réseau à diffusion et garantissant la cohérence d'un système de numérotation des paquets global et réparti**
Annick Le Coz, Hervé Le Goff, Michel Ollivier , 31 pages ; Octobre 1982
- PI 182 **Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence**
Israël César Lerman , 34 pages ; Novembre 1982
- PI 183 **L'IRISA vu à travers les stages effectués par ses étudiants de DEA (1^{ère} année de thèse)**
Daniel Herman , 41 pages ; Novembre 1982
- PI 184 **Commande non linéaire robuste des robots manipulateurs**
Claude Samson , 52 pages ; Janvier 1983
- PI 185 **Dialogue et représentation des informations dans un système de messagerie intelligent**
Philippe Besnard, René Quiniou, Patrice Quinton, Patrick Saint-Dizier, Jacques Siroux, Laurent Trilling , 45 pages ; Janvier 1983
- PI 186 **Analyse classificatoire d'une correspondance multiple ; typologie et régression**
I.C. Lerman , 54 pages ; Janvier 1983
- PI 187 **Estimation de mouvement dans une sequence d'images de télévision en vue d'un codage avec compensation de mouvement**
Claude Labit , 132 pages ; Janvier 1983
- PI 188 **Conception et réalisation d'un logiciel de saisie et restitution de cartes élémentaires**
Eric Sécher , 45 pages ; Janvier 1983
- PI 189 → *sur le point de paraitre*
- PI 190 **Généralisation de l'analyse des correspondances à la comparaison de tableaux de fréquence**
Brigitte Escofier , 35 pages ; Mars 1983
- PI 191 **Association entre variables qualitatives ordinales «nettes» ou «floues»**
Israël-César Lerman , 42 pages ; Mars 1983
- PI 192 } *sur le point de paraitre*
- PI 193 }
- PI 194 **Régime stationnaire pour une file M/H/1 avec impatience**
Raymond Marie et Jean Pellaumail , 8 pages ; Mars 1983

62

63

64

65

66

67