REVUE DE STATISTIQUE APPLIQUÉE

J. PAGÈS

# Analyse factorielle de données mixtes

*Revue de statistique appliquée*, tome 52, n° 4 (2004), p. 93-111

<http://www.numdam.org/item?id=RSA_2004__52_4_93_0>

*Rev. Statistique Appliquée,* 2004, **LII** (4), 93-111

# ANALYSE FACTORIELLE DE DONNÉES MIXTES

J. PAGÈS

*Laboratoire de mathématiques appliquées*
*Pôle d'Enseignement Supérieur et de Recherche Agronomique, 35042 Rennes cedex*
*email : pages@agrorennes.educagri.fr*

## RÉSUMÉ

Une méthodologie factorielle permettant d'inclure à la fois des variables quantitatives et qualitatives en tant qu'éléments actifs d'une même analyse a été proposée par B. Escofier (1979) dans le cadre de l'ACM. Il est montré ici que cette approche se confond d'une part avec la méthodologie esquissée par Saporta (1990a) dans le cadre de l'ACP, et d'autre part avec une AFM dans laquelle chaque variable constitue un groupe à elle seule. L'ensemble de ces trois points de vue confère à la méthode proposée par B. Escofier le statut d'une méthode à part entière dotée de plusieurs bonnes propriétés et facile à mettre en œuvre. En combinant cette méthode avec l'AFM, il est possible de réaliser une AFM sur des groupes de variables mixtes ce qui est une possibilité nouvelle. Une application de cette extension de l'AFM est présentée.

***Mots-clés :*** *Analyse en composantes principales, Analyse des correspondances multiples, Analyse factorielle multiple, Données mixtes.*

## ABSTRACT

In the framework of multiple correspondence analysis, B. Escofier (1979) proposed a factor analysis in which both quantitative and qualitative variables can intervene as active ones. Here it is shown that this approach is equivalent to the one outlined by Saporta (1990a) in the principal components analysis framework and to a multiple factor analysis (MFA) in which each set of variables is composed by only one variable. All these equivalences lead to a method (factor analysis for mixed data : FAMD) having several good properties and easy to perform. Combining FAMD and MFA, it is possible to extend the field of MFA to mixed sets of variables. An application of such an extension of MFA is described.

***Keywords :*** *Principal components analysis, Multiple correspondence analysis, Multiple factor analysis, Mixed data.*

L'introduction simultanée de variables quantitatives et qualitatives (données dites mixtes) en tant qu'éléments actifs d'une même analyse factorielle est une problématique fréquente. La méthodologie usuelle consiste à transformer les variables quantitatives en qualitatives en découpant en classes leur intervalle de variation et à soumettre le tableau homogène ainsi obtenu à une analyse des correspondances multiples (ACM). Cette méthodologie est relativement facile à mettre en œuvre et

éprouvée dès lors que les individus sont un tant soit peu nombreux, disons au delà d'une centaine pour fixer les idées, limite en deçà de laquelle l'ACM donne des résultats peu stables.

L'intérêt de conserver telles quelles les variables quantitatives reprend ses droits dans deux cas :

- lorsque le nombre de variables qualitatives est très petit comparé à celui des variables quantitatives : ainsi, peut-on hésiter à recoder vingt variables quantitatives dans le seul but de pouvoir introduire une seule variable qualitative;
- lorsque le nombre d'individus est faible.

Plusieurs propositions d'analyse factorielle de données mixtes ont déjà été faites : on en trouvera une synthèse concise dans Saporta (1990a). Beaucoup de travaux reposent sur la notion de «codage optimal» c'est-à-dire l'affectation, satisfaisant un certain critère, de valeurs numériques aux modalités. Ainsi codées, les variables qualitatives peuvent être traitées comme des variables quantitatives (Young, 1981; Tenenhaus & Young, 1985) ce qui conduit, lorsque les variables sont toutes qualitatives, à des méthodologies dont l'intérêt n'est pas évident pour un utilisateur de l'ACM. Mais ceci ouvre naturellement la voie à un traitement simultanée des deux types de variables (Tenenhaus, 1977; Young *et al.*, 1978). Adoptant le point de vue de l'ACM, Escofier (1979) a proposé d'introduire des variables quantitatives (moyennant un codage approprié) dans une ACM : elle décrit plusieurs propriétés de cette méthodologie ainsi qu'une application.

Il est possible, moyennant une métrique judicieusement choisie, de réaliser une ACP sur un tableau juxtaposant des variables quantitatives réduites et des variables qualitatives codées sous forme disjonctive complète. Cette possibilité est esquissée dans Saporta (1990a) sous le nom d'extension de l'ACP et de l'ACM mais l'auteur, après avoir mentionné les insuffisances des représentations des variables auxquelles cette extension conduit, ne conclut pas à un intérêt pratique de la méthode.

Enfin, lorsque les variables constituent des groupes homogènes (*i.e.* les variables d'un même groupe sont de même type), une analyse factorielle multiple (AFM) peut être réalisée (Escofier & Pagès, 1998, p. 173; Pagès, 2002).

Dans cet article, nous reprenons les idées de B. Escofier (1979) et les transposons dans le cadre de l'ACP. On obtient ainsi une nouvelle possibilité de mettre en œuvre la méthode qu'elle proposa, possibilité qui se confond avec l'extension citée par Saporta (1990a). En regroupant les propriétés induites par les deux points de vue, on obtient une méthode assez riche que nous appelons Analyse Factorielle de Données Mixtes (AFDM). L'AFDM est ensuite comparée à la pratique empirique qui consiste à introduire directement des indicatrices dans une ACP. On montre enfin son équivalence avec une AFM dans laquelle chaque groupe est réduit à une variable, quantitative ou qualitative.

En conjuguant la variante ACP de la méthode de B. Escofier et l'AFM, on obtient une méthode factorielle prenant en compte des groupes de variables mixtes, c'est-à-dire pouvant inclure des variables des deux types. Les possibilités de l'AFM sont ainsi enrichies. La mise en œuvre de cette extension peut-être réalisée avec un programme usuel d'AFM. Un exemple illustre cette possibilité.

## 1. Données, notations

Nous disposons de $I$ individus. Usuellement, chaque individu $i$ est muni du poids $p_i$ tels que $\Sigma_i p_i = 1$. Pour simplifier, nous supposons les individus de même poids, soit $p_i = 1/I \; \forall i$. Ces individus sont décrits par :

- $K_1$ variables quantitatives $\{k = 1, K_1\}$ ; ces variables seront toujours supposées centrées-réduites; ceci n'est pas une commodité mais une nécessité due à la présence des deux types de variables;
- $Q$ variables qualitatives $\{q = 1, Q\}$; la $q^e$ variable présente $K_q$ modalités $\{k_q = 1, K_q\}$ ; l'ensemble des modalités a pour cardinal $\Sigma_q K_q = K_2$ ; on note $p_{k_q}$ la proportion des individus possédant la modalité $k_q$.

Soit $K = K_1 + K_2$ le nombre total de variables quantitatives et de variables indicatrices.

Ces notations peuvent être rassemblées dans le tableau de la figure 1 dans lequel les variables qualitatives apparaissent à la fois sous leur forme condensée et sous leur forme disjonctive complète.

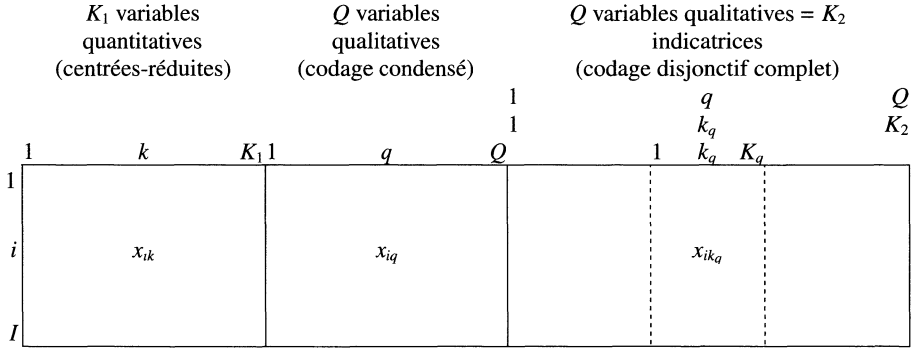


FIGURE 1
*Structure des données et principales notations*
$x_{ik}$ *: valeur de i pour la variable (centrée-réduite) k*
$x_{iq}$ *: modalité de i pour la variable q*
$x_{ik_q} := 1$ *si i possède la modalité* $k_q$ *de la variable q et 0 sinon*

## 2. Représentation des variables dans $R^I$

Soit $R^I$ l'espace des fonctions sur $I$. Cet espace est muni de la métrique diagonale des poids des individus notée $D$ : $D(i,j) = 0$ si $j \neq i$

$= p_i$ si $j = i$

Généralement les individus ont le même poids : $D = (1/I)\boldsymbol{I_d}$ (en notant $\boldsymbol{I_d}$ la matrice identité de dimension convenable).

Comme en ACP normée, les variables quantitatives sont représentées par des vecteurs de longueur 1.

Comme en ACM, la variable $q$ est représentée par le nuage $N_q$ de ses $K_q$ indicatrices centrées. Ce nuage engendre le sous-espace $E_q$ de dimension $K_q - 1$, ensemble des fonctions sur $I$ centrées et constantes sur les classes de la partition définie par $q$. Pour que $N_q$ ait, dans une ACP non normée, les mêmes propriétés inertielles que dans une ACM, il faut affecter à l'indicatrice $k_q$ le poids $1/p_{k_q}$ (en toute rigueur, obtenir exactement l'inertie de l'ACM nécessite le poids $1/Qp_{k_q}$, ce qui « moyenne » les inerties par le nombre de variables, propriété indésirable ici où les variables qualitatives sont confrontées à des variables quantitatives dont les inerties ne sont pas « moyennées »). Comme les programmes d'ACP usuels ne permettent pas l'introduction directe de poids de colonnes, on préfèrera diviser les valeurs de l'indicatrice $k_q$ par $\sqrt{p_{k_q}}$. On parlera alors de *codage* ACM de l'indicatrice $k_q$.

En procédant ainsi, on obtient en particulier la propriété fondamentale suivante de l'ACM : l'inertie projetée de $N_q$ sur une variable centrée $y$ est égale au carré du rapport de corrélation $\eta^2(q, y)$ entre $q$ et $y$. En effet, en notant $\overline{y}_{k_q}$ la moyenne de $y$ pour les individus possédant la modalité $k_q$ et $s_y^2$ la variance de $y$, on a, puisque la colonne $k_q$ codée ACM a pour moyenne $\sqrt{p_{k_q}}$ et que $y$ est centré :

$$\sum_{k_q \in q} \left\langle k_q, \frac{y}{\|y\|} \right\rangle^2 = \sum_{k_q \in q} \left( \sum_i \left[ \frac{1}{I} \frac{x_{ik_q} - p_{k_q}}{\sqrt{p_{k_q}}} \cdot \frac{y_i}{s_y} \right] \right)^2 = \frac{\sum_{k_q \in q} p_{k_q} \overline{y}_{k_q}^2}{s_y^2} = \eta^2(q, y)$$

Si l'on choisit $y$ dans $E_q$, on obtient $\eta^2(q, y) = 1$ : dans $E_q$, l'inertie de $N_q$ est isotrope, propriété classique de l'ACM.

En recherchant la direction $v$ de $R^I$ qui rend maximum l'inertie projetée du nuage $N_K$ (comportant les variables quantitatives et les indicatrices), on rend maximum le critère :

$$\sum_{k \in K_1} r^2(k, v) + \sum_{q \in Q} \eta^2(q, v)$$

point de départ de la proposition de Saporta (1990a p.66).

Géométriquement, les variables $k$ étant réduites, $r(k, v) = \cos\theta_{kv}$, en notant $\theta_{kv}$ l'angle entre les vecteurs $k$ et $v$. De même, $v$ étant centrée, $\eta^2(q, v) = \cos^2\theta_{qv}$ en notant $\theta_{qv}$ l'angle entre $v$ et sa projection sur $E_q$ (cette propriété est généralement établie avant le centrage des indicatrices, *cf.* Saporta 1990b p.149; elle vaut aussi après ce centrage). Le critère s'écrit alors

$$\sum_{k \in K_1} \cos^2\theta_{kv} + \sum_{q \in Q} \cos^2\theta_{qv}$$

point de départ de la présentation de l'AFDM par Escofier (1979).

Dans sa présentation de l'AFDM, Escofier (1979) adopte un point de vue technique symétrique de celui choisi ici : elle se place dans le cadre de l'ACM et

code la variable quantitative de façon à obtenir un tableau traitable dans ce cadre. Il s'agit donc bien de la même méthode, dont les résultats peuvent être obtenus de différentes façons.

*Remarque sur l'influence des deux types de variables*

Dans l'espace $R^I$ :

- une variable quantitative est représenté par un vecteur associé à une inertie de 1 ;
- une variable qualitative à $K_q$ modalités est représenté par $K_q$ vecteurs engendrant un sous-espace $E_q$ de dimension $K_q - 1$, l'ensemble étant associé à une inertie de $K_q - 1$.

Comme en ACM, l'inertie totale d'une variable qualitative est d'autant plus grande qu'elle présente beaucoup de modalités; mais en projection sur une direction quelconque de $E_q$, cette inertie vaut 1. En ce sens, les variables des deux types sont équilibrées dans la recherche de directions d'inertie maximum, ce qui est bien traduit par l'une ou l'autre des deux écritures du critère ci-dessus.

## 3. Représentation des individus dans $R^K$

L'espace $R^K$ a pour dimensions les $K_1$ variables quantitatives et les $K_2$ indicatrices. Il est muni de la métrique euclidienne usuelle.

La distance entre les individus $i$ et $l$ s'écrit :

$$d^2(i,l) = \sum_{k \in K_1} (x_{ik} - x_{lk})^2 + \sum_{q \in Q} \sum_{k \in K_q} \frac{1}{p_{k_q}} (x_{ik_q} - x_{lk_q})^2$$

avec comme cas particulier important, la distance entre un individu et le centre de gravité du nuage. Ce centre de gravité est confondu avec l'origine $O$ dès lors que les variables sont centrées ce que nous avons supposé d'emblée pour les variables quantitatives; pour les indicatrices codées ACM, compte tenu de la division par $\sqrt{p_{k_q}}$, la moyenne de la colonne $k_q$ vaut $\sqrt{p_{k_q}}$.

On obtient finalement :

$$d^2(i,O) = \sum_{k \in K_1} x_{ik}^2 + \sum_{q \in Q} \sum_{k \in K_q} \left( \frac{x_{ik_q}}{\sqrt{p_{k_q}}} - \sqrt{p_{k_q}} \right)^2 = \sum_{k \in K_1} x_{ik}^2 + \sum_{q \in Q} \frac{1 - p_{q(i)}}{p_{q(i)}}$$

en notant $q(i)$ la modalité de la variable $q$ possédée par $i$ et $p_{q(i)}$ la proportion associée.

Les variables quantitatives contribuent à cette distance exactement comme elles le font dans l'ACP portant sur ces seules variables; les variables qualitatives contribuent à cette distance (au coefficient $1/Q$ près) comme elles le font dans l'ACM de ces seules variables.

Il reste à s'assurer de l'équilibre entre les influences des deux types de variables dans ces relations. Il est naturel de mesurer l'influence d'une variable par

sa contribution à l'inertie de l'ensemble des points. Les considérations établies dans $R^I$ se transposent dans $R^K$ par dualité. En particulier, dans le sous-espace de $R^K$ engendré par les $K_q$ modalités de la variable $q$, la projection du nuage des individus a une inertie de $K_q - 1$ répartie de façon isotrope dans toutes les directions de ce sous-espace.

## 4. Graphiques

Comme dans toute analyse factorielle on représente :

- le nuage des individus par sa projection sur ses axes d'inertie (on note $F_s(i)$ la projection de l'individu $i$ sur l'axe de rang $s$);
- les variables quantitatives par leur coefficient de corrélation avec les facteurs $F_s$;
- les modalités de variables qualitatives par les centres de gravité des individus correspondant (on note $F_s(k_q)$ la projection, sur l'axe de rang $s$, du centre de gravité des individus possédant la modalité $k$ de la variable $q$).

Il peut être commode de représenter les indicatrices comme des variables quantitatives. Le coefficient de corrélation entre une indicatrice $k_q$ et le facteur $F_s$ peut s'écrire, en utilisant le fait que le facteur $F_s$ est centré (Dagnélie, 1998, tome 1, p. 133) :

$$r(k_q, F_s) = \sqrt{\frac{p_{k_q}}{1 - p_{k_q}}} \frac{F_s(k_q)}{\sqrt{\lambda_s}}$$

Il est égal au rapport de corrélation entre $F_s$ et la variable qualitative à deux modalités associée à $k_q$. Par rapport à la représentation des centres de gravité, on écarte les points correspondant à un effectif fort.

## 5. Relations de transition

On applique ici les formules générales de l'ACP au tableau codé comme indiqué en 2.

### *5.1. Relations de $R^I$ vers $R^K$*

Soit $G_s(k)$ la coordonnée de la colonne $k$ sur l'axe de rang $s$.

Cas d'une variable quantitative :

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{1}{I} x_{ik} F_s(i) = r(k, F_s)$$

Cas d'une modalité $k_q$ de la variable $q$ ayant la fréquence relative $p_{k_q}$. On a, puisque $F_s$ est centré :

$$G_s(k_q) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{1}{I} \left( \frac{x_{ik_q}}{\sqrt{p_{k_q}}} - \sqrt{p_{k_q}} \right) F_s(i)$$

$$= \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{1}{I} \frac{x_{ik_q}}{\sqrt{p_{k_q}}} F_s(i) = \frac{\sqrt{p_{k_q}}}{\sqrt{\lambda_s}} F_s(k_q)$$

La représentation des indicatrices transformées $(G_s(k_q))$ ne semble pas présenter d'avantages par rapport à celle des centres de gravité $(F_s(k_q))$.

### 5.2. *Relation de transition de $R^K$ dans $R^I$*

Cette relation est fondamentale en ACM où elle exprime la position d'un individu par rapport aux modalités qu'il possède. Elle est rarement explicitée en ACP mais est sous-jacente aux interprétations. Pour l'AFDM, elle s'écrit :

$$F_s(i) = \frac{I}{\sqrt{\lambda_s}} \sum_{k \in K_1} x_{ik} G_s(k) + \frac{1}{\sqrt{\lambda_s}} \sum_{k_q \in K_2} \left( \frac{x_{ik_q}}{\sqrt{p_{k_q}}} - \sqrt{p_{k_q}} \right) G_s(k_q)$$

Le premier membre est celui de l'ACP usuelle. Il exprime qu'un individu se trouve globalement du côté des variables pour lesquelles il a une valeur au-dessus de la moyenne et à l'opposé des variables pour lesquelles il a une valeur au-dessous de la moyenne.

Le second peut s'écrire en fonction de $F_s(k_q)$, grâce à la seconde relation du §5.1 :

$$\frac{1}{\lambda_s} \sum_{k_q \in K_2} \left( x_{ik_q} - p_{k_q} \right) F_s(k_q) = \frac{1}{\lambda_s} \sum_{k_q \in K_2} x_{ik_q} F_s(k_q)$$

Il exprime qu'un individu est, au coefficient $\lambda_s$ près, au barycentre des modalités qu'il possède. Finalement, un individu se trouve à la fois du côté des variables pour lesquelles il a une forte valeur et du côté des modalités qu'il possède.

## 6. Comparaison avec l'ACP avec indicatrices (ACPi)

On peut songer à introduire directement des variables qualitatives codées en disjonctif complet dans une ACP (normée), technique que nous notons ACPi. Une comparaison entre l'ACPi et l'AFDM peut être réalisée rapidement au travers du nuage des indicatrices qu'une même variable $V$, présentant $m$ modalités, induit (dans $R^I$) dans les deux cas.

Ces indicatrices engendrent le même sous-espace $E_V$ à $(m-1)$ dimensions dans les deux méthodes. Leur inertie totale vaut $m$ (ACPi) ou $m-1$ (AFDM) : ces inerties diffèrent donc surtout pour $m$ faible.

La répartition de cette inertie dans $E_V$ est isotrope dans l'AFDM et irrégulière dans l'ACPi. On peut approcher intuitivement cette répartition au travers du coefficient de corrélation $r(a,b)$ entre les indicatrices de deux modalités $a$ et $b$ d'une même variable. En notant $p_a$ (resp. $p_b$) la fréquence relative de la modalité $a$ (resp. $b$), on a :

$$r(a,b) = -\sqrt{\frac{p_a p_b}{(1-p_a)(1-p_b)}}$$

Si toutes modalités ont le même effectif, $r(a,b) = -1/(m-1)$. Cette valeur constante, combinée avec la dimension $m-1$ de $E_V$, assure l'isotropie de l'inertie dans $E_V$, propriété fondamentale de l'ACM (démonstration en annexe). Il en résulte que, lorsque les modalités d'une même variable possèdent le même effectif :

- si en outre les $Q$ variables présentent le même nombre $m$ de modalités, l'ACPi conduit, à un coefficient près, aux mêmes valeurs propres et aux mêmes composantes principales que l'ACM. Dans ce cas, l'inertie totale des nuages analysés vaut $Qm$ en ACPi et $m-1$ en ACM : on passe donc des valeurs propres de l'ACPi à celle de l'ACM par le coefficient $(m-1)/Qm$;
- si les variables ne possèdent pas le même nombre de modalités, l'ACPi équivaut à une ACM pondérée (Cazes, 1980; Pagès 2002) particulière dans laquelle chaque variable (ayant $m$ modalités) est, par rapport à l'ACM, surpondérée par $m/(m-1)$. En pratique, rien ne justifie cette pondération.

En ACPi, la coordonnée de la modalité $a$ le long de l'axe $s$ (notée $G_s(a)$) est égale au coefficient de corrélation entre l'indicatrice de $a$ et la composante principale $v_s$ noté $r(v_s,a)$. Ce coefficient est relié à la coordonnée (le long de l'axe de rang $s$) du centre de gravité des individus possédant $a$ (noté $F_s(a)$) par la relation :

$$G_s(a) = r(a,v_s) = \sqrt{\frac{p_a}{1-p_a}}\frac{F_s(a)}{\sqrt{\lambda_s}}$$

Dans le cas ci-dessus de l'équivalence entre ACPi et ACM (lorsque toutes les modalités ont le même effectif), la représentation des modalités est identiques à celle de l'ACM (au coefficient $\sqrt{1/(m-1)}$ près). Sinon, par rapport à la représentation de l'ACM, les modalités sont d'autant plus éloignées de l'origine qu'elles possèdent un effectif important.

Lorsque les effectifs des modalités d'une même variable diffèrent, le relation plus haut exprimant $r(a,b)$ montre que les modalités de faible effectif sont proches de l'orthogonalité et les modalités de fort effectif ont tendance à s'opposer. Ceci peut être illustré géométriquement dans le cas de variables à 3 modalités puisque dans ce cas ces modalités appartiennent à un plan de $R^I$ (figure 2).

La modalité $c$ est fixée sur l'axe horizontal. Les secteurs en pointillés montrent les domaines de variation de $a$ et $b$ selon leur effectif. Il en résulte que l'ACPi fait jouer, dans la construction des premiers axes un rôle plus grand aux modalités ayant un fort effectif.

*Conclusion.* L'ACP sur tableau disjonctif complet (ACPi) n'est évidemment pas une pratique à recommander. Elle peut néanmoins conduire à des résultats empiriquement

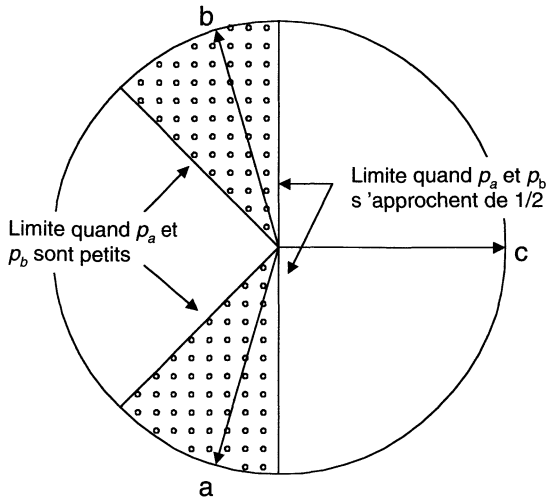


FIGURE 2
*Représentation de l'ensemble $\{a, b, c\}$ des indicatrices des trois modalités d'une variable qualitatives dans le sous-espace (de $R^I$) qu'elles engendrent. La modalité c est fixée sur l'axe horizontal; les secteurs pointillés montrent les domaines de variation de a et b selon leurs effectifs*

exploitables du fait de son équivalence avec l'ACM lorsque toutes les modalités ont le même effectif quelle que soit la variable.

## 7. Equivalence avec l'analyse factorielle multiple (AFM)

En AFM, les groupes de variables sont pondérés de façon à rendre égale à 1 leur inertie axiale maximum. En introduisant un tableau de données mixtes dans lequel chaque variable, quantitative ou qualitative, constitue un groupe, on obtient donc les résultats de l'AFDM :

- les variables quantitatives sont centrées-réduites;
- les variables qualitatives sont codées comme en ACM.

L'idée d'appliquer l'AFM à des groupes constitués chacun d'une seule variable quantitative ou qualitative à déjà été proposée (Abascal-Fernandez *et al.* 2003).

## 8. Analyse Factorielle Multiple pour groupes de Données Mixtes

L'AFM usuelle permet de traiter simultanément des groupes de variables quantitatifs ou qualitatifs (Escofier et Pagès 1998; Pagès 2002). En combinant l'AFDM et l'AFM, il est possible d'étendre l'AFM au cas de groupes de variables pouvant inclure chacun des variables des deux types. Ces groupes sont codées de façon à ce que leur ACP non normée conduise aux résultats de l'AFDM; dans l'AFM, ils

sont alors déclarés comme quantitatifs. En procédant ainsi, on équilibre à la fois les groupes entre eux et les variables au sein de chaque groupe. On retrouve ici un cas particulier d'analyse factorielle multiple hiérarchique (Le Dien et Pagès 2003).

## 9. Mise en œuvre concrète

*AFDM à partir d'un programme d'AFC (Escofier 1979)*

C'est la technique originelle. Les variables qualitatives sont introduites selon le codage disjonctif complet. Chaque variable quantitative $k$ est codée sur deux colonnes ayant pour l'individu $i$ respectivement les valeurs $(1 - x_{ik})/2$ et $(1 + x_{ik})/2$.

*AFDM à partir d'un programme usuel d'ACP*

Les variables quantitatives doivent être au préalable centrées et réduites puisque l'on utilise l'ACP non normée. Les variables qualitatives apparaissent au travers de leurs indicatrices dans laquelle $x_{ik_q}$ $(= 0$ ou $1)$ est divisé par $\sqrt{p_{k_q}}$ (il n'est pas utile de centrer puisque les programmes usuels d'ACP le font).

L'ACP fournit directement les représentations des individus et des variables quantitatives. Pour obtenir la représentation des centres de gravité des modalités, on introduit aussi les variables qualitatives déclarées comme telles en supplémentaire (lorsque cela est possible, cas du logiciel SPAD par exemple).

*AFDM à partir d'un programme d'AFM*

Les variables quantitatives sont introduites brutes et les variables qualitatives sous leur forme condensée. Chaque variable constitue un groupe.

*AFM sur groupes mixtes*

Les groupes mixtes sont codés comme pour une AFDM via une ACP et sont déclarés comme quantitatifs non réduits. Les autres groupes sont codés comme usuellement.

## 10. Application

### *10.1. Données, problématique*

La méthodologie proposée ici sera appliquée aux données de Russet, popularisées en France par Tenenhaus (1998, 1999), qui les emprunta à Gifi (1990), pour illustrer l'approche PLS. Ces données comportent plusieurs groupes de variables mais un seul de ces groupes sera pris en compte ici.

On dispose pour 47 pays de quatre mesures de l'instabilité politique que Tenenhaus décrit ainsi :

- La variable INST est une fonction du nombre de responsables du pouvoir exécutif et du nombre d'années pendant lesquelles le pays a été indépendant entre 1945 et 1961. Cet indice varie entre 0 (très stable) et 17 (très instable).

- La variable ECKS est l'indice d'Eckstein calculé sur la période 1946-1961. Il mesure le nombre de conflits violents entre communautés sur cette période.
- La variable DEAT est le nombre de personnes tuées lors de manifestations violentes sur la période 1950-1962.
- La variable DEMO classe les pays en trois groupes : démocratie stable, démocratie instable et dictature.

Dans son application, Tenenhaus utilise la transformation des données introduites par GIFI (transformation que nous adopterons aussi) : Exp (Inst-16,3), Ln (Ecks + 1) et Ln (Deat + 1).

La variable Demo est introduite au travers de ses indicatrices : *démocratie stable, démocratie instable, dictature.*

L'objet de l'application est de comparer les deux codages : d'une part l'introduction des indicatrices comme des variables quantitatives usuelles comme le fait Tenenhaus, c'est à dire centrées-réduites et, d'autre part, l'introduction des indicatrices codées ACM comme indiqué au §2. Concrètement, les données se présentent sous la forme d'un tableau rectangulaire (*cf.* Figure 3) ayant 47 lignes (les pays) et deux blocs de 6 colonnes comprenant chacun les mêmes variables (les trois variables quantitatives auxquelles s'ajoutent les trois indicatrices) mais codées différemment. Le premier bloc, noté $A$, contient les variables brutes et les indicatrices usuelles (0 ou 1) : il sera centré-réduit dans l'analyse. Le second, noté $B$, contient les variables quantitatives centrées-réduites et les indicatrices codées ACM : dans l'analyse, il ne sera que centré.

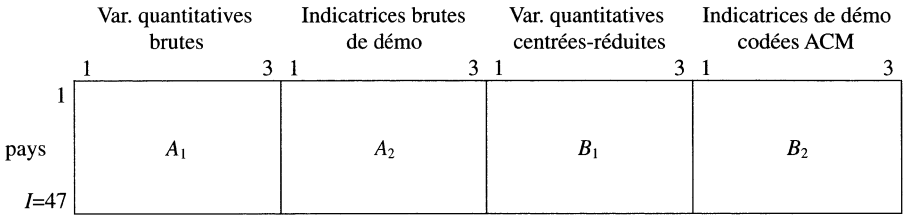


FIGURE 3
*Structure du fichier soumis à l'AFM*

*Dans l'analyse, le tableau* $A = [A_1|A_2]$ *sera réduit tandis que le tableau* $B = [B_1|B_2]$ *ne le sera pas; d'où la réduction préalable de* $B_1$.

### *10.2. Méthodologie*

Pour comparer les résultats de l'ACP du tableau $A$ d'une part et du tableau $B$ d'autre part (cette seconde ACP est équivalente à une AFDM), une AFM dans laquelle chacun de ces deux tableaux constitue un groupe actif est réalisée. Ainsi, on illustre en outre l'AFM de groupes mixtes.

Dans cette analyse, le groupe A est centré-réduit; le groupe B est seulement centré, d'où le centrage-réduction préalable des variables quantitatives. Ainsi, d'un groupe à un autre : les variables quantitatives sont exactement les mêmes; les

indicatrices engendrent le même sous-espace mais avec des contributions à l'inertie totale (et une répartition de cette inertie) différentes.

Ce cas particulier ne comportant qu'une seule variable qualitative, la contribution à l'inertie des indicatrices dans le tableau $B$ est particulièrement simple : nombre d'indicatrices $-1$.

### *10.3. Analyses séparées*

TABLEAU 1
*Inerties respectives des deux tableaux*

| | | Décomposition de l'inertie totale | | | | | Décomposition de l'inertie des 3 indicatrices | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Inertie totale | par type de variables | | par axe (en%) | | | dans l'analyse globale | | | dans leur analyse séparée | |
| | | quanti. | quali. | $F_1$ | $F_2$ | $F_3$ | $F_1$ | $F_2$ | $F_1$+$F_2$ (%) | $F_1$ | $F_2$ |
| ACP avec ind. | 6 | 3 | 3 | 45.3 | 27.7 | 12.8 | 1.26 | 1.27 | 58% | 1.609 | 1.391 |
| AFDM | 5 | 3 | 2 | 46.0 | 26.0 | 13.6 | .66 | .67 | 38% | 1 | 1 |

*$F_1 + F_2$ (%) : % de l'inertie du premier plan dû aux indicatrices. Ex :* $.58 = (1.26 + 1.27)/[6(.453 + .277)]$

Du fait du faible nombre de modalités, l'influence a priori des indicatrices (inertie totale) est sensiblement plus élevée dans l'ACPi (*cf.* Tableau 1).

Du fait des effectifs voisins des modalités, (15, 12, 20), la répartition de l'inertie dans le sous-espace qu'elles engendrent est assez régulière dans l'ACPi (elle varie au maximum de 1.609 à 1.391). La différence entre les résultats des deux analyses devrait provenir essentiellement de l'influence relative de la variable qualitative.

TABLEAU 2
*Corrélation entre les facteurs des analyses séparées*

| | | AFDM | | |
|---|---|---|---|---|
| | | *F*1 | *F*2 | *F*3 |
| ACPi | *F*1 | .99 | -.08 | .03 |
| (avec | *F*2 | .07 | .97 | .24 |
| indicatrices) | *F*3 | -.04 | -.23 | .97 |

*$-.08$ : coefficient de corrélation entre le premier facteur de l'ACPi et le second de l'AFDM.*

Le tableau 2 montre une grande parenté entre les deux analyses. Toutefois les analyses ne coïncident pas : leur comparaison fine se fera à l'aide de l'AFM.

### *10.4 Indicateurs globaux de l'AFM*

Comme attendu, l'AFM met en évidence la grande parenté entre les deux analyses. Ses deux premiers facteurs sont très proches de leurs homologues dans les analyses séparées.

TABLEAU 3
*Quelques indicateurs globaux de l'AFM*

| | inertie totale | | inertie par groupe | | Corrélation avec les facteurs des analyses séparées | | | |
|---|---|---|---|---|---|---|---|---|
| | brute | en % | *ACPi* | *AFDM* | *ACPi F1* | *ACPi F2* | *AFDM F1* | *AFDM F2* |
| *AFM F1* | 1.993 | 45.5% | 1.00 | 1.00 | .997 | .042 | .998 | .034 |
| *AFM F2* | 1.168 | 26.7% | .61 | .56 | .045 | .993 | .031 | .991 |

### *10.5 Interprétation générale du premier plan de l'AFM (figures 4 et 5)*

Le premier plan fait apparaître les 3 pôles correspondant aux modalités de la variable qualitative. Le nombre de conflits violents est faible dans les démocraties stables; le nombre de tués est plus élevé dans les dictatures; le nombre de gouvernements est élevé dans les démocraties instables.

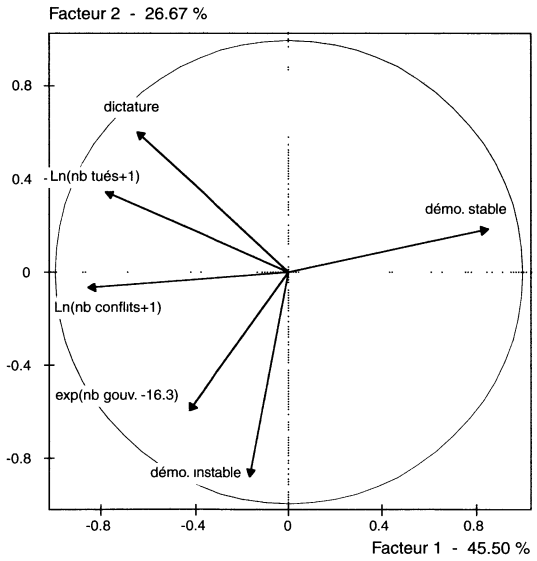


FIGURE 4
*Représentation des variables quantitatives et des indicatrices*
*Les deux codages des indicatrices conduisent aux mêmes coefficients de corrélation avec les facteurs*

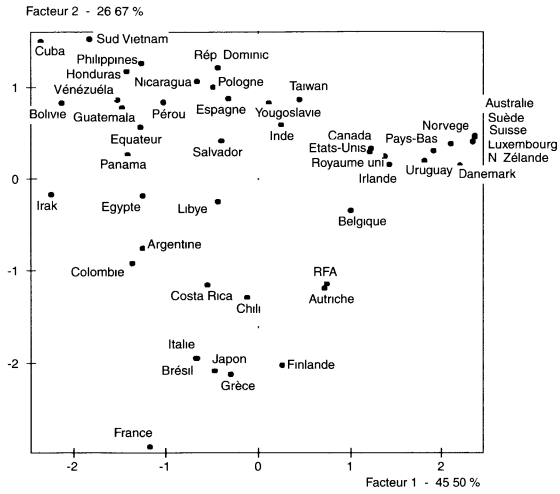


FIGURE 5
*Représentation des individus*

Du point de vue des pays, par exemple, la Suède est typique des démocraties stables, la France typique des démocraties instables et Cuba typique des dictatures. Ces interprétations se retrouvent facilement dans les données (*cf.* tableau 4).

TABLEAU 4
*Données des quelques pays commentés dans le texte.*
*Les variables quantitatives sont centrées-réduites*

| pays | régime polit. | nb gouvernements | nb conflits violents | nb tués |
|---|---|---|---|---|
| Argentine | démo. instable | –0,573 | 1,129 | 1,505 |
| Cuba | dictature | –0,573 | 1,541 | 2,630 |
| France | démo. instable | 3,169 | 0,973 | –0,534 |
| Inde | démo. stable | –0,843 | 1,404 | 0,342 |
| Libye | dictature | 1,590 | –0,253 | –0,835 |
| Suède | démo. stable | –0,841 | –1,883 | –0,835 |
| Taiwan | dictature | –0,843 | –0,854 | –0,835 |

### *10.6. Représentation superposée des nuages partiels*

Tous les pays ont leurs deux images partielles proches l'une de l'autre, conséquence de la forte structure commune entre les deux analyses. Les plus fortes différences (Taiwan et Inde pour l'axe 1; Libye et Argentine pour l'axe 2) sont représentées figure 6.

La différence attendue entre les 2 méthodologies est la suivante : le rôle de la variable qualitative est plus important dans l'ACPi (inertie totale : 3) que dans

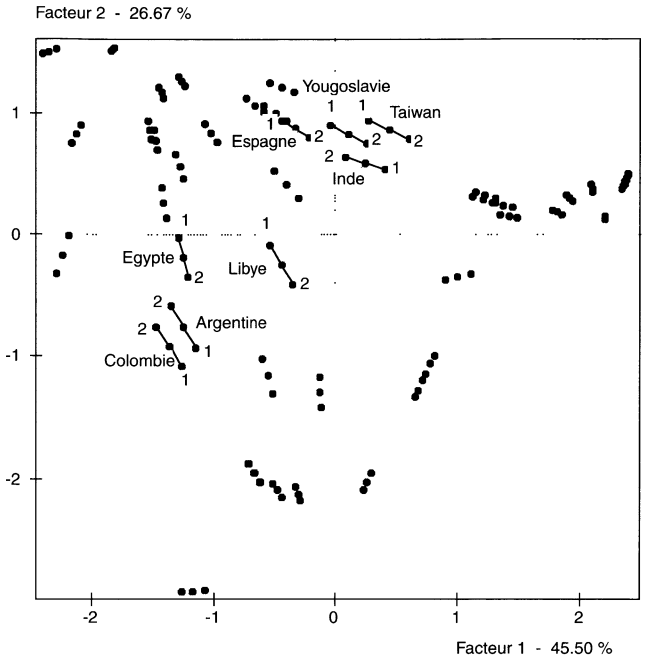


FIGURE 6
*AFM : représentation des individus moyens et partiels. Seuls les pays commentés dans le texte sont identifiés. 1 : ACPi; 2 : AFDM*

l'AFDM (inertie totale : 2). Concrètement, dans les données soumises à l'AFM, cela se traduit par des valeurs différentes pour une même indicatrice d'un bloc à l'autre : l'indicatrice $k$ a une variance de 1 dans le bloc associé à l'ACPi et $(1 - p_k)$ dans le bloc associé à l'AFDM. Ainsi, dans l'AFM, un individu partiel du groupe ACPi aura ses coordonnées plus influencées par la variable qualitative que son homologue du groupe AFDM.

Les plus forts écarts entre points partiels homologues le long de l'axe 1 sont observés pour l'Inde et Taiwan. Ces deux pays occupent une position intermédiaire entre les pôles *dictature* et *démocratie stable* car ils possèdent chacun des caractéristiques des deux pôles : Taiwan est classée *dictature* mais présente les caractéristiques d'une *démocratie stable* (faibles valeurs des trois indicateurs quantitatifs; *cf.* tableau 4) : par rapport au point partiel Taiwan – AFDM, le point Taiwan – ACPi est plus proche du pôle dictature du fait de la plus grande influence de la variable qualitative dans l'ACPi. L'Inde possède des caractéristiques presque opposées : elle est classée *démocratie stable* mais présente un nombre de tués et surtout un nombre de conflits violents élevé.

Le long de l'axe 2, les plus forts écarts entre individus partiels homologues sont observés pour la Libye et l'Argentine. Ces deux pays occupent une position intermédiaire entre les deux pôles *dictature* et *démocratie instable* car ils possèdent chacun des caractéristiques de ces deux pôles : la Libye, classée *dictature*, a eu un nombre très élevé de gouvernements pendant la période 1945- 1961; au contraire

l'Argentine, classée *démocratie instable*, a eu un nombre peu élevé de gouvernements pendant la période 1945-1961. A titre d'exemple, pour ces deux pays, on peut tenir un raisonnement analogue à celui illustré à propos de l'Inde et de Taiwan. Le point Libye-ACPi est plus proche du pôle *dictature* que le point Libye-AFDM, du fait du plus grand rôle joué par la variable *régime politique* dans l'ACPi.

Dans cet exemple on peut relier aisément la représentation superposée de l'AFM et celles des analyses séparées. En effet, les deux analyses séparées, conduisent à des facteurs très voisins de ceux de l'AFM (le coefficient de corrélation entre un facteur de l'AFM et le facteur de même rang de l'une quelconque des deux analyses séparées est toujours supérieur à .99; *cf.* tableau 3).

Ainsi pour les huit pays dont les images partielles sont représentées, la configuration sur le plan de l'AFM des images partielles d'un groupe se retrouve presque parfaitement sur le plan issu de l'analyse séparée de ce groupe (*cf.* figures 7 et 8). Cet exemple illustre bien, dans un cas de figure nouveau (présence d'un groupe mixte), ce que l'on attend de la représentation superposée issue de l'AFM : des indications sur la structure des nuages associés à chacun des groupes.

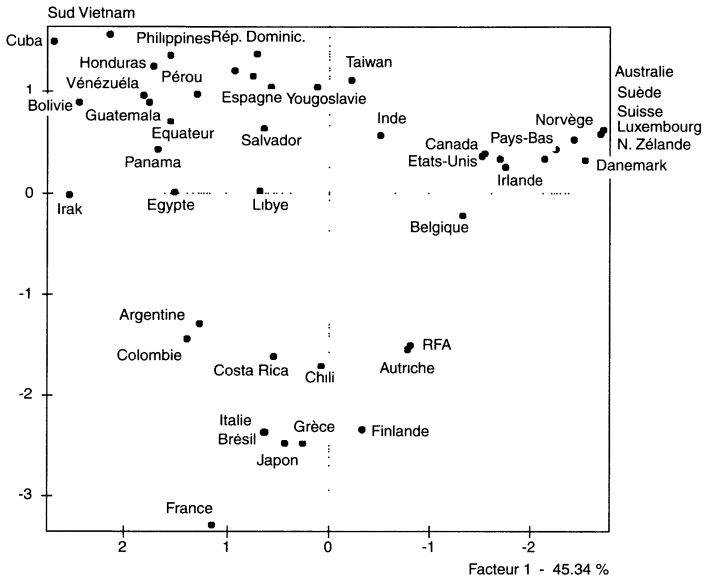


FIGURE 7
*ACPi : représentation des individus (pays)*

### *10.7. Conclusion quant à l'exemple*

Cet exemple montre la grande parenté entre les résultats issus de l'ACPi et de l'AFDM dans un cas où cette parenté est attendue : une seule variable qualitative, des effectifs peu différents d'une modalité à l'autre.

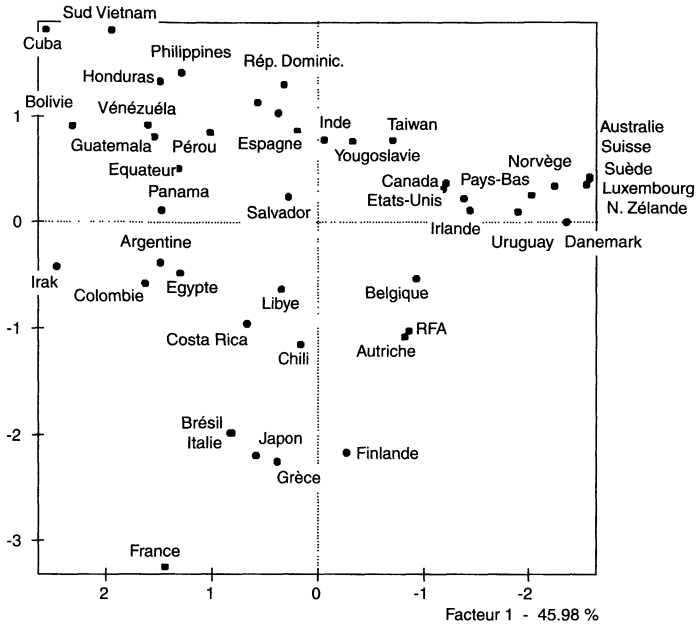


FIGURE 8
*AFDM : représentation des individus (pays)*

Les plus grande différences entre ces deux analyses concernent des individus présentant des caractéristiques opposées selon que l'on considère les variables quantitatives ou qualitatives : c'est pour ces individus que la différence de pondération des variables qualitatives entre les deux analyses se fait le plus sentir.

La présence de groupes incluant les deux types de variables dans une AFM ne semble pas engendrer d'artefacts et conduit à des graphiques interprétables avec les règles usuelles.

## 11. Conclusion

La méthodologie proposée initialement par Escofier (1979) pour introduire simultanément des variables quantitatives et qualitatives dans une analyse factorielle présente suffisamment de bonnes propriétés et de potentiel d'application pour justifier le statut d'une méthode à part entière : l'Analyse Factorielle de Données Mixtes (AFDM).

Elle prend en compte les variables quantitatives comme une ACP normée et les variables qualitatives comme une ACM. L'équilibre entre les deux types de variables est assuré au sens de l'inertie axiale maximum de chaque variable.

Les résultats qu'elle produit peuvent être interprétés avec les règles usuelles de l'ACP et de l'ACM.

Sa mise en œuvre peut être réalisée facilement avec un programme d'ACP ou d'AFC et très facilement à l'aide d'un programme d'AFM.

En combinant l'AFDM et l'AFM, on peut étendre l'AFM au cas de groupes mixtes : une telle AFM traite les groupes quantitatifs comme une ACP (normée ou non) les groupes qualitatifs comme une ACM, et les groupes mixtes comme une AFDM.

## Annexe

### Structure de l'inertie associée à une variable qualitative dans l'ACPi lorsque ses $m$ modalités sont équiprobables (*cf.* §6).

On adopte le point de vue de l'ACPi des modalités d'une seule variable. La matrice à diagonaliser, notée $C$, comporte des 1 sur la diagonale et $a = -1/(m-1)$ partout ailleurs.

On a : $C = (1-a)\boldsymbol{I_d} + a\mathbf{1}\mathbf{1}'$

$\mathbf{1}$ étant le vecteur dont les $m$ composantes sont égales à 1 et $\boldsymbol{I_d}$ la matrice identité. On a alors :

$$C\mathbf{1} = (1-a)\mathbf{1} + a\mathbf{1}\mathbf{1}'\mathbf{1} = [1+(m-1)a]\mathbf{1} = 0$$

Ainsi, $\mathbf{1}$ est vecteur propre de $C$ relatif à la valeur propre 0.

Si $u$ est un vecteur perpendiculaire à $\mathbf{1}$ : $\mathbf{1}'u = 0$

$$Cu = (1-a)u + a\mathbf{1}\mathbf{1}'u = (1-a)u = [m/(m-1)]u$$

Ainsi, l'hyperplan orthogonal à $\mathbf{1}$ est un espace propre associé à la valeur propre $[m/(m-1)]$

Il en résulte que, pour une variable ayant une distribution uniforme, l'inertie du nuage de ses $m$ modalités est répartie de façon isotrope dans un sous-espace de dimension $m-1$.

## Références

ABASCAL-FERNANDEZ E., LANDALUCE-CLUO M.I., GARCIA-LAUBE I. (2003), Multiple factor analysis of mixed tables : a proposal for analysing problematic metric variables. *Proceeding cf CARME 2003 meeting*. Barcelona, june 2003.

CAZES P. (1980), L'analyse de certains tableaux rectangulaires décomposés en blocs. *Les cahiers de l'analyse des données*, 5 (1) 9-23 et 5 (2) 133-154.

DAGNELIE P. (1998), *Statistique théorique et appliquée*. De Boeck Université Ed.

ESCOFIER B.(1979), Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. *Les cahiers de l'analyse des données*, 4 (2) 137-146.

ESCOFIER B. et PAGES J. (1998), *Analyses factorielles simples et multiples*. 3e ed. Dunod.

GIFI A. (1990), *Non linear multivariate analysis*. John Wiley & Sons, Chichester.

LE DIEN S. ET PAGES J. (2003), Analyse factorielle multiple hiérarchique. *Revue de statistique appliquée*, LI (2), 47-73.

PAGES J. (2002), Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de statistique appliquée*, L (4), 5-37.

SAPORTA G. (1990a), Simultaneous analysis of qualitative and quantitative data. *Atti della XXXV riunione scientifica; società italiana di statistica*, 63-72.

SAPORTA G. (1990b), *Probabilités, analyse des données et statistique*. Technip, Paris.

SPAD (2002), Diffusé par DECISIA, 30 rue Victor Hugo, 92532 Levallois-Perret cedex.

TENENHAUS M. (1977), Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée*, XXV (2), 39-56.

TENENHAUS M. (1998), *Régression PLS : Théorie et Pratique*. Technip, Paris.

TENENHAUS M. (1999), «L'approche PLS», *Revue de Statistique Appliquée*, 47, 2, 5-40.

TENENHAUS M., YOUNG F. (1985), An analysis and synthesis of multiple correspondance analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50 (1), 91-119.

YOUNG F. W. (1981), Quantitative analysis of qualitative data. *Psychometrika*, 46 (4), 357-388.

YOUNG F. W., TAKANE Y., DE LEEUW J. (1978), The principal components of mixed measurements level multivariate data : an alternating least squares method with optimal scaling features. *Psychometrika*, 43 (2), 279-281.