BUREAU D'APPLICATION DES METHODES
STATISTIQUES ET INFORMATIQUES

BAMSI REPRINT 04/2003

Introduction à l'analyse des données

Samuel AMBAPOUR

BAMSI B.P. 13734 Brazzaville
BAMSI REPRINT 04/2003

Introduction à l'analyse des données(*)

Samuel AMBAPOUR(**)

Ce cahier n'est pas un cours.

On y insiste sur le traitement pratique des données et sur les applications des différentes méthodes d'analyse. Un même exemple illustratif est utilisé tout au long de l'exposé et sert de base pour la comparaison des méthodes utilisées.

Pour des exposés théoriques complets de ces méthodes, le lecteur est invité à consulter les ouvrages de base cités en référence.

Grâce à l'outil informatique et notamment à de nombreux logiciels commercialisés sur micro-ordinateurs, l'utilisateur de l'analyse des données peut désormais se consacrer aux tâches essentielles à savoir, le choix de la méthode et l'interprétation des résultats.

Dans ce cahier, il est fait usage du logiciel ADDAD diffusé par l'association du même nom(***) (''Association pour le Développement et la Diffusion de l'Analyse des Données'').

(*) Ce texte a été publié dans ''les cahiers du CASP'' n°3-4, décembre 1992
(**) Enseignant au CASP
(***) Ce cahier s'inspire, au niveau de la forme et du langage, des travaux de cette association.

TABLE DES MATIERES

''Avec l'Analyse des Données fondée sur l'usage de l'ordinateur, c'est une nouvelle méthodologie que la statistique apporte à la science et notamment aux sciences de l'homme''.

J-P. Benzécri

‘’L’Analyse des Données n’est certes pas simplement un ensemble de techniques nouvelles et, sans être le vecteur philosophique de la recherche du sens de toute chose, c’est quand même une nouvelle manière d’être, face à un tableau de données”.

J-P. Fenelon.

…”Les services rendus montrent bien que l’Analyse des Données constitue aujourd’hui, et de loin, la partie la plus immédiatement rentable de la statistique”.

G. Morlat

1. Introduction

Il n’y a pas très longtemps, on ne pouvait pas traiter un tableau de 3000 lignes et 300 colonnes. L’apparition et le développement des ordinateurs a du coup levé cet obstacle de calcul, et a permis la conservation et l’exploitation des grandes masses de données. Cette amélioration continue de l’outil informatique a fortement contribué au développement et à la vulgarisation de nombreuses méthodes statistiques, devenues maintenant d’usage assez courant.

Aujourd’hui, des vastes données d’enquêtes sont dépouillées et, fournissent de grands tableaux qui se prêtent aisément à l’interprétation. Des données issues d’investigations spécifiques sont rassemblées et constituent une masse importante et apparemment indéchiffrable d’informations mais, qu’on peut désormais traiter sans difficultés.

Cependant, comment ‘’extraire les phénomènes, les lois, les connaissances que recèlent ces données que nous ne pouvons appréhender directement’’[8] ?

La statistique classique nous a habitué à étudier les variables les unes après les autres, de construire autant d’histogrammes que de variables. Comment faire pour que, à ces

nombreux graphiques se substitue un seul graphique, une carte plane ? Comment devant, la profusion des descriptions parcellaires fournies par l'analyse variable par variable, donner une vision globale de l'ensemble des résultats ? Les techniques dites *d'analyse des données* permettent de répondre à ces questions.

Pour J-P. Fénelon ''l'analyse des données est un ensemble de techniques pour découvrir la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de traduire par une structure plus simple et qui la résume au mieux. Cette structure peut le plus souvent, être représentée graphiquement''[ ] 31 .

Ces techniques qui sont essentiellement descriptives, ont pour but de décrire, de réduire, de classer et de clarifier les données en tenant compte de nombreux points de vue et d'étudier, en dégageant les grands traits, les liaisons, les ressemblances ou les différences entre les variables ou groupes de variables. Les documents fournis sont qualifiés de ''synthétiques et percutants et valent souvent mieux qu'un long discours''. Cette approche descriptive et multidimensionnelle permet de dire que l'Analyse des Données, c'est de la ''statistique descriptive perfectionnée''.

L'analyse des données recouvre principalement deux ensembles de techniques : ''les premières qui relèvent de la géométrie euclidienne et conduisent à l'extraction de valeurs et de vecteurs propres, sont appelées *''analyses factorielles''* ; les secondes, dites de *''classification automatique''* sont caractérisées par le choix d'un indice de proximité et d'un algorithme d'agrégation ou de désagrégation qui permettent d'obtenir une partition ou arbre de classification''[53].

Parmi ces deux techniques, les premières occupent une place de choix, ''car elles sont utilisées soit seules, soit conjointement avec les secondes, alors que ces dernières sont rarement appliquées seules''[28].

On s'intéressera surtout aux analyses factorielles dont on ne décrira que les deux méthodes les plus employées. Il s'agit de *l'analyse en composantes principales* (beaucoup utilisée dans les pays anglo-saxons) et de *l'analyse factorielle des correspondances* (très prisée en France). La classification automatique sera introduite comme aide à l'interprétation d'une analyse factorielle. Ce qui permet de compléter et d'enrichir les résultats de cette dernière. Cependant, vu la diversité des méthodes, on

regardera comment se présentent les résultats pour l'une d'entre elles : *la classification*

ascendante hiérarchique, qui est la plus élaborée des méthodes de classification.

2. Un peu d’histoire

Bien que l'étude de la structure de vastes ensembles de données soit récente, les principes dont les méthodes d'analyse de données s'inspirent sont anciens.

En ce qui concerne l'analyse factorielle, il faut remonter aux travaux de Ch. Spearman (1904) qui introduit pour la première fois le concept de facteur ; il cherche, derrière les notes obtenues par de nombreux sujets à de nombreux tests, une variable explicative cachée : le facteur général d'aptitude (analyse factorielle au sens des psychologues).

C'est vers les années 30 que se pose le problème de la recherche de plusieurs facteurs (travaux de C. Burt et de L.L Thurstone) ; on cherche deux puis plusieurs facteurs : mémoire, intelligence, etc. ''non observables directement mais susceptibles d'expliquer au sens statistique du terme les nombreuses notes obtenues par les sujets''. Comme on le constate il s'agissait déjà de résumer à l'aide d'un petit nombre de facteurs une information multidimensionnelle. De nos jours on ne fait guère appel à l'analyse factorielle au sens des psychologues parce qu'elle suppose un modèle a priori.

Puis, l'analyse factorielle en composantes principales développée par H. Hotelling (1933), mais dont on peut faire remonter le principe à K. Pearson (1901) : ''les individus colonnes du tableau à analyser étant considérés comme des vecteurs d'un espace à p dimensions, on proposait de réduire la dimension de l'espace en projetant le nuage des points individus sur le sous-espace de dimension k (k petit fixé) permettant d'ajuster au mieux le nuage''[53]. D'un point de vue plus récent écrit L. Lebart, l'analyse au composantes principales est «une technique de représentation des données, ayant un caractère optimal selon certains critères algébriques et géométriques spécifiés et que l'on utilise en général sans référence à des hypothèses de nature statistique ou à un modèle particulier''[43].

Enfin, l'analyse factorielle des correspondances introduite par J.P Benzécri (1962), est actuellement en vogue. Elle fournit, sans hypothèses a priori des représentations simplifiées dans un certain sens à l'interprétation. Laissons sur ce point la parole au Professeur J.P Bensécri : ''l'analyse des correspondances telle qu'on la pratique en 1977 ne se borne pas à extraire des facteurs de tout tableau de nombres positifs. Elle donne pour la préparation des données des règles telles que le *codage sous-forme disjonctive complète* ; aide à critiquer la validité des résultats, principalement par des calculs de

contribution ; fournit des procédés efficaces de discrimination et de régression ; se conjugue harmonieusement avec la classification automatique''[6]. Sa logique est claire : le modèle doit suivre les données non l'inverse ; le modèle probabiliste est jugé trop contraignant : ''statistique n'est pas probabilité''.

Les deux méthodes précédentes et celles qui en ont été dérivées, comme l'analyse *factorielle discriminante* (initiée par Fisher en 1936, qui permet de décrire la liaison entre une variable qualitative et un ensemble de variables quantitatives) et l'analyse *canonique* (introduite par Hotelling en 1936 et dont l'objectif initial était d'exprimer au mieux à l'aide d'un petit nombre de couples de variables la liaison entre deux ensembles de caractères quantitatifs) dépendent d'un même corps de résultats mathématiques qu'on exposera dans le paragraphe ''analyse générale''.

S'agissant de la classification automatique, compte tenu de ''la multiplicité des techniques existantes et l'effervescence qui règne autour de ce domaine'', car selon R.M. Cormack (cité par Lebart) plus de 1000 articles sont publiés par an sur ce thème, il est vraiment difficile de faire l'historique de ces méthodes ; en effet nombreux sont les chercheurs qui ont contribué à leur mise en œuvre et dont les précurseurs sont : Buffon
(1749), Adanson (1757) et Linné (1758). ''Je me contenterai de rapprocher les objets, suivant le plus grand nombre de degrés de leurs rapports et leur de leurs ressemblances… Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore ensemble afin d'en faire un tout dont les parties soient unies et liées intimement'' écrivait Adanson''[47].

Pour terminer cette page d'histoire, mentionnons l'analyse des données non métriques introduite par une nouvelle école de statisticiens américains sous le nom de « *multidimensional scaling* » (J.D. Carrol, J.B. Kruskal, R.N. Shepard, …) et dont les principales méthodes sont :

- l'analyse des proximités ;
- l'analyse des préférences ;
- l'analyse de mesure conjointe (qui permet d'expliquer une variable qualitative ordinale à l'aide des variables nominales).

Ces méthodes ont trouvé leurs applications surtout dans le domaine du

marketing[9]. 10

3. Types de tableaux analysables

Les données se présentent généralement sous la forme d'un tableau rectangulaire, dont les lignes correspondent à des individus ou unités statistiques et les colonnes à des variables appelées caractères ou caractéristiques.

Les valeurs des variables peuvent être :

- quantitatives ordinales (jugement humain, température) ;
- quantitatives mesurables (poids d'un individu, revenu) ;
- qualitatives ordinales (classe d'âge, le rang) ;
- qualitatives nominales (sexe, situation matrimoniale).

Lorsque dans un tableau, toutes les variables choisies sont quantitatives, on peut établir un tableau de données quantitatives ; c'est le cas par exemple où l'on observe sur un ensemble de sujets I , un certain nombre de mesures J : poids, taille, âge. Ce tableau est encore appelé *tableau de mesures*.

A partir de deux variables qualitatives, on peut définir un *tableau de contingence* croisant les modalités de deux variables, l'ensemble des lignes correspond aux modalités de la première variable et l'ensemble des colonnes aux modalités de la deuxième variable ; par exemple le tableau qui répartit la population congolaise recensée en 1974 selon les deux caractères ‘’région’’ et ‘’classe d'âge’’.

Si l'on divise chaque valeur du tableau précédent par le cardinal de la population, on obtient le tableau de fréquences relatives que l'on appellera simplement *tableau de fréquence*.

Si l'on croise plus de deux variables qualitatives entre elles définies sur une même population, on peut construire un tableau contenant l'ensemble des tableaux de contingence entre les variables prises deux à deux. Le tableau ainsi obtenu est appelé *tableau de Burt*. C'est un tableau symétrique qui comporte sur sa diagonale ''des résultats qu'en terme de dépouillement d'enquête on appellerait des ''tris à plats'', alors qu'ailleurs on a tous les tableaux des ''tris croisés'' des variables deux à deux.

On rencontre aussi des tableaux de *préférence*. Un ensemble I d'individus donne des J jugements de préférence globale sur un ensemble d'objets ; on demande par exemple à chaque personne interrogée de noter de 1 à 4 l'ordre de préférence pour quatre marques de bière : primus, kronenbourg, ngok, amstel. A l'intersection de la $i\,è - me$ ligne et de la $j - ème$ colonne, on trouve le rang attribué par la personne i à la bière j.

Le tableau de préférence est différent du tableau de *rang.* Reprenons le tableau de contingence qui répartit la population congolaise selon les deux caractères ''région'' et ''classe d'âge''. On obtient un tableau de rang si à l'intersection de la région i et de la classe d'âge j, on y inscrit le rang de la région sur toutes les régions, relativement à l'effectif de la classe d'âge i. Dans le tableau de préférence rencontré ci-haut, la ligne j est une permutation de 4 objets alors que dans le tableau de rang c'est la colonne qui est une permutation de nombres de 1 à 9 (les 9 régions du Congo).

Les tableaux de *proximités* évoluent la similarité ou la dissimilarité entre chaque couple d'individus par un indice de proximité ou de distance (tableau de distance inter-villes).

Souvent, on observe des variables qui ne prennent que deux valeurs codées généralement 0 et 1 ; elles conduisent à des tableaux *binaires* : par exemple un individu doit répondre par ''oui'' ou par ''non'' à une question ; le ''oui'' est codé 1, le ''non'' est codé 0 ; on peut aussi citer le cas des *tableaux de présence-absence* où il s'agit du relevé de la présence ou de l'absence d'un caractère. Tel ménage possède ou ne possède pas le caractère : avoir un poste téléviseur : la présence est codé 1, l'absence est codé 0.

D'une manière générale, un tableau rempli uniquement de 0 et de 1 est appelé *tableau logique*. C'est le cas des tableaux précédents. Nous verrons au § 6.4.1, qu'on peut transformer un tableau de données quantitatives en un tableau de description logique par découpage en classes des variables quantitatives. En fait, ''parler de tableau logique, c'est désigner un certain format de codage, qui peut recouvrir des domaines très [3] différen . ts'' 1

On peut également mentionner *les tableaux de notes*. Il s'agit dans le cas qui nous intéresse des notes scolaires (type de tableaux analysé dans ce cahier) comprises entre deux bornes (0 et 20). Ce tableau peut être analysé comme tel (c'est ce que nous ferons dans les chapitres suivants). Dans bien de cas, pour donner la même importance à chaque observation, on ''dédoublera'' chaque colonne du tableau, c'est-à-dire qu'à

12

chaque matière d'origine on lui fait correspondre une matière dite ''duale'' : avoir 15/20 en statistique, c'est avoir 5/20 en la matière duale. L'analyse factorielle d'un tableau de notes dédoublé semble d'un point de vue pratique donner des résultats plus clairs et plus facilement interprétables que l'analyse du tableau initial[12]. Le tableau de description logique décrit précédemment peut être considéré comme un tableau de notes particulier dans lequel toutes les notes ne peuvent prendre que l'une des valeurs 0 ou 1.

Pour terminer, on peut citer les tableaux de *correspondance chronologique* ou tableaux ternaires ou encore tableaux multiples. C'est par exemple le cas du tableau où, I est J l'ensemble d'industries (ou produits), un ensemble de pays, T un ensemble d'époques, désignant les échanges pour le produit , à l'instant t en provenance

$_{IJT}k\ i$

(ou à destination) du pays . Une généralisation au cas quaternaire a été étudiée et on j obtient un tableau de la forme où

$_{IJPT}k\ I$ est par exemple l'ensemble des pays

J

exportateurs, l'ensemble des mêmes pays considérés comme exportateurs, P est un ensemble des classes de produits et un ensemble d'époques : $_{IJPT}k$ est donc la valeur

T

des importations du pays i en provenance du pay j (ou des exportations du pays à s j destination du pays i , rentrant dans la classe de produits) p t effectuées en l'année t . e

P our l'étude de ces types de tableaux, on utilise très largement la technique des ''points supplémentaires'' (cf §5.2.5)[ ] 14 .

Le tableau soumis à l'analyse doit posséder certaines qualités : *pertinence, homogénéité, exhaustivité.* Il ne faut retenir dans la masse hétérogène des faits que ce qui se rapporte à un seul point de vue (pertinence), et ne pas mélanger les quantités exprimées en kilogrammes et en mètres (homogénéité). L'exhaustivité implique que les différentes zones du domaine d'investigation sont bien représentées. A ces trois exigences ''il faut ajouter une condition assez évidente, mais parfois oubliée : le tableau de données doit être vaste et en statistique, l'infini est parfois de l'ordre de 30"[42].

4. Analyse générale

On part d'un tableau rectangulaire reliant deux ensembles finis I et . On a Ca $J\ rdI$ observations sur lesquelles sont mesurées $CardJ$ variables : $_{ij}$ x est la mesure de l a i $i\ I$. $(CardI = n, CardJ = p)$. $_{ij}x$ peut être la note variable j de J sur l'indiv du de obtenue par l'étudiant i à l'épreuve j .

Le tableau X peut admettre deux représentations [35] :

- l'une dans un espace vectoriel R^n avec un nuage de points correspondant chacun à p une ligne ;

- l'autre dans un espace vectoriel R^p avec un nuage de n points correspondant chacun à une colonne.

L'analyse factorielle revient à faire la recherche des *axes principaux d'inertie* (ou axes factoriels) des deux nuages. On cherche donc à ajuster le nuage des points par un sous-espace vectoriel de R^p, muni de la distance euclidienne usuelle (c'est-à-dire que le carré de la distance entre deux points est égal à la somme des carrés des différences de leurs coordonnées). On commence par déterminer une droite passant par l'origine et ajustant au mieux le nuage à étudier, en minimisant la somme des carrés des distances des points à la droite. Ce calcul conduit à un vecteur unitaire porté par cette droite dit aussi vecteur propre relatif à une valeur propre. De façon analogue on peut continuer l'ajustement et trouver dans R^p un certain nombre de vecteurs propres et de valeurs propres toutes positives décroissant avec le rang. X étant la matrice du tableau, et $'X$ la matrice transposée, u_α les vecteurs propres et λ_α les valeurs propres seront solutions de l'équation :

$'XXu\ u_\alpha = \lambda_\alpha\alpha$ dans R^p

Le vecteur u est norme par la relation :

$'u\ u = 1$

Le premier axe factoriel est donc le vecteur $_1u$ correspondant λ_1 la plus grande valeur propre de $'XX$. L'inertie expliquée par cet axe est λ_1.

En prolongeant le problème on trouve que le sous-espace qui explique la plus grande inertie contient les premiers vecteurs propres $_1,\ldots,_q u\ u$ e $'q$ d XX. L'inertie expliquée par ce sous-espace est égale à la somme des valeurs propres correspondant à ces vecteurs propres. On aura les formules correspondantes dans R^n. En effet, il est démontré que [43] :

- si v_α est vecteur propre unitaire de $'XX$ relatif à la valeur propre $0\ \lambda_\alpha \neq$, $^-$ = t vecteur unitaire de $'XX$ es relatif à la même valeur propre.

$^{1/2}'u\ X\ v_\alpha\ \lambda_\alpha\alpha$

-réciproquement, si u_α est vecteur unitaire de $'XX$ relatif à $0\ \lambda_\alpha \neq$, $^{1/2}v\ Xu_\alpha\ \lambda_\alpha\alpha$ $^-$ est = vecteur unitaire de $'XX$ relatif à λ_α.

u_α est appelé $\alpha -$ *ème axe factoriel* dans ^pR.

v_α est appelé $\alpha -$ *ème axe factoriel* dans ^nR.

5. Analyse en composantes principales

5.1. Les données – les objectifs

En analyse en composantes principales, l'ensemble I est décrit à l'aide de variables p

quantitatives, continues, homogènes ou non a priori corrélées entre elles deux à deux. On cherche à répondre à des questions du type suivant : quelles sont les variables qui sont liées positivement entre elles ? Quelles sont celles qui s'opposent ? A propos des individus on cherchera à évaluer leur ressemblance et leur dissemblance, à mettre en relief des groupes homogènes d'individus. En résumé l'analyse en composantes principales (ACP) consiste à transformer les p variables quantitatives, initiales en p nouvelles variables non corrélées, appelées composantes principales (ou facteurs).

5.2. La méthode

On ne décrira ici, que l'une des variantes de cette méthode et qui est de loin la plus employée : *l'analyse en composantes principales normées.* On suppose que les données de départ sont non seulement hétérogènes quant à leur moyenne, mais le sont également quant à leur dispersion et à leur nature (disparité des unités de mesure). Pour ramener chaque variable à un cadre commun de comparabilité, on opérera sur chacune d'elle une transformation linéaire ramenant sa moyenne à zéro et sa variance à l'unité (variable centrée réduite).

5.2.1. Le tableau de données

On a les mêmes ensembles I et J de l'analyse générale.

$X = \in \{x\, i\, _{ij} I, j \in J\}$

$_{j} x$

.
.
.

i …….. $_{ij} x$ ……

.
.
.
.

On calcule :

i)- La moyenne de la variable $_{j} x$ ***:***

$$\lceil \rceil \blacksquare \rfloor \sum \in \text{avec } M = \in \sum \{m\, i\, _{i} I\}$$

$$\blacksquare\, i\, I \qquad m \qquad \}$$

$$_{j\, ij} \qquad i$$

$$M$$

où est le poids affecté à l'individu ; $m_{i\,i}\,1$

$$m^{i}\ CardI = \text{et}\ \{\}\ 1\ m\ i_{i} \sum \in I = = M$$

ii)- La variance de la variable $_{j}x$ **:**

$$\sigma^{\lceil\rceil} = = \quad \underset{x\,x\,x\,i\,I}{\overset{m}{\overset{-\in}{\lfloor\rfloor\sum}}}\ (\ )^{2\,2}\,\mathrm{var}(\ )_{j\,ij\,j}$$

$$\overset{j}{M}$$

iii) I **:**

La variable centrée et réduite qui a pour composantes sur l'ensemble $x -$

$$x = \quad \overset{ij\,j}{X\sigma}$$

$$\overset{ij}{\qquad\qquad}$$

où $\overset{j}{\qquad}$ type de $_{j}x$

σ_{j} est l'écart

$$(\ )\{\}\ 0\ mo_{j\,ij} \Rightarrow {}^{y\,X = \in \sum X\,i\,I =} \qquad \sum\{\} = \in \lfloor\rfloor$$

$$\overset{i}{\overset{ij}{\overset{jj\,ij}{''}}}$$

et $\{(\ )\}^{2}$

$$\in = \sum \lfloor\rfloor \qquad M$$

$$\mathrm{var}(\ )\ 1\ XX_{j\,ij}m_{j} = -\lceil\rceil\ oy\ X\,i\,I$$

iv)- Le coefficient de corrélation

linéaire entre deux variables ^{m} r [redacted] $_{i}x$ ***et*** $_{j}x$ **,:**

$\lceil\rceil$

leurs entre –1 et +1.

qui prend les va

5.2.2. Analyse des points individus i de $(\ )\,N\,I_{J}$ dans ^{p}R

On se placera au centre de gravité du nuage des points de base (normalisation centrée réduite). Le $i\,è - me$ individu sera représenté dans l'espace des variables normées X_{j} par un point ayant pour coordonnée la valeur X_{ij} et affecté de masse (poids) . Si l'on

$$m_{i}$$

note par :

$$^{m}\ N\,I\,X\,i\,I\lceil\rceil$$

= ∈ ⌊ ⌋ { } le nuage des points $i I \in$;

( ) (; ^{i}

$J ij$

M

i) Le centre de gravité G ***de ce nuage a pour*** j *è* − *me* ***coordonnée*** **:** ^{m} X

$i I$

$$= \in \{ \} \lfloor \rfloor \sum_{i} \quad GJ\,ij$$

$$M \quad \lceil \rceil \lceil \rceil \quad m\,m \quad i\,I$$

$$= {}^{1}_{0}{}^{i\,i}$$

$$\sigma_{ij\,j} \quad \lceil \rceil \in = \lfloor \sum \sum ; \lfloor \lfloor \rfloor \lfloor \rfloor \rfloor \quad {}_{j}M\,M$$

c’est donc l’origine du système d’axes dans lequel est placé le nuage des individ

us.

$( ) N I {}^{p}_{j} R$ **:**

ii) La distance entre deux points de s’écrit dans $\{ \}^{2}$ $d\,i($

$$,i)^{2}( )^{ij}_{i\,j} = - \sum X\,Xj \in J$$

$$\lceil \rceil - \lfloor \rfloor = \in \{ \lfloor \rfloor \quad \sum_{x\,x_{j}\,J}$$

$$^{2}(\quad {}^{ij}_{i\,j}$$

σ ^{j}

(c

’est la distance euclidienne ‘’usuelle’’). Ainsi chaque variable aura une contribution égale à la dispersion totale du nuage $( ) N I$

J .

vaut **:** 17

$( ) N I_{j}$ ***iii) La distance d’un point de*** i ***au centre de gravité G du nuage***

$$\{ \}^{222}(,) ( )_{ij} \quad d\,i\,G = = \rho\,i \sum Xj \in J$$

iv) L’inertie d’un point i ***par rapport au centre de gravité est*** **:** ^{m} $I\,I\,i$

$^{2}( ) ( )^{i}$

n

$M = \rho$

et l’inertie du nuage ( ) sera égal à : $N I_{J}$

$$m_{I N I}\,i\,I$$

$$= \in \{\} \sum$$

$$^{2}(\,(\,))\,{}^{i}$$

$$_{n J}$$

$$M$$

$$\{\} \in m\, i\, I$$

$$\sum \sum = {}^{2}\, {}^{i}_{ij}$$

$$M \in$$

$$m\, i\, I\, j\, J$$

$$\in \in \sum\sum$$

$$= {}^{2}, {}^{i}_{ij}$$

$$M$$

$$= \sum\{Var(\,)\,X_{j} j \in J\}$$

or $\mathrm{var}(\,)\,1\,X_{j} =$

$(\,(\,))_{n J} \Rightarrow = I N I = CardJ$

L'inertie du nuage des points i est donc égale au nombre de variables ; cette inertie est aussi égale à la somme des termes diagonaux (trace) de la matrice de corrélation entre les variables dont le terme général est $'_{jj}$

r . C'est donc cette matrice qu'il faudra diagonaliser pour la re valeurs propres.
cherche des vecteurs et

v) Les facteurs et axes factoriels-Coordonnées des observations dans l'espace factoriel.

Soient $\{F\,i_{\alpha}(\,)\,i \in I\}$ composantes principales normées.
les facteurs associés à l'analyse en

Les facteur e moyenne nulle, de variance égale à λ_{α} , et sont deux orthogonaux.
s sont d

En effet :

$$\,_{0}\, m^{i}\, i\, I_{\alpha}$$

$$\{\}$$

$$\in = \{$$

$$M$$

$$^{2}\, m^{i}\, i\, I$$

= ∈ { } ⌊ ⌋ ∑

$\lambda_{\alpha\alpha}$

M

et

[redacted] 0 ⌈m^i [redacted] $i\ I$

M ⌉ { } $_{\alpha\beta}$ ∈ = ⌊ ⌋ ∑ si ⌉α ≠ β

On sait déjà que la somme de toutes les valeurs propres est égale au nombre $CardJ$ de variables. Et donc :

(( )) $_{nJ}$ $I\ N\ I\ CardJ$ = = $\sum \lambda_{\alpha}$

5.2.3. Analyse des points variables j de ( ) NJ_I dans ^nR

En ACP, l'origine des axes n'est pas le centre de gravité du nuage des variables ; les axes factoriels issus du nuage des individus ne sont pas les axes principaux d'inertie du

$^2 d\ j\ (0,\ )$ =1

nuage des variables. On a vu que ( ) 1 $Var\ X_j$ = c'est-à-dire que ; les variables X_j sont donc situées sur une sphère de rayon 1 centrée en 0, origine initiale des axes.

L'intersection de la sphère et d'un plan factoriel est un cercle dit *cercle de corrélation.*

La distance euclidienne usuelle entre deux points de ( ) NJ_I dans ^{n}

R : (,) ( ) $_{ij\ ij}$ $d\ j\ j$ = − ∑ $XX\ i \in I$

{ } $^{2\ '2}$

En tenan it ( ) 1 ∑ $XX\ r$ =

$_j$

t compte du fa que $Var(\ )\ X_j$ = $Var\ X$, = et , On trouve que : $_{ij\ jj}$

$^{2\ '}$(,) 2(1) $_{jj}$

$d\ j\ j$ = − r ,

r , st le coeffi ient de corrélation linéaire entre les variables j et $'j$. Ainsi, les où , e c

$_{jj}$

proximités entre points variables s'expriment en termes de corrélations :

$j\ 'r\ 1\ j$

= ⇒ les points et sont confondus ; $'_{jj}$

r = −1 ⇒ les points j et $'j$ sont diamét alem nt op o

r e p sés sur la sphère (0,1) ;

$'_{jj}$

r = ⇒ les points j et $'j$ sont orthogonaux et se trouvent aux extrémités d'un arc de

$_{jj}$

, 0

90°. qui existent entre les matrices

5.2 $(\,)\,N\,I_{J}j\,(\,)\,N\,J_{I}$

.4- Relation entre les points de et de

Nous avons vu au chapitre 4 les relations $^{'}X\,X$ et $^{'}XX$ en

ce qui concerne les vecteurs et les valeurs propres. En utilisant ces propriétés, on peut

J

établir les relations de transition entre les facteurs $F\,i(\,)_{\alpha}$ de I et $G\,j\,(\,)_{\alpha}$ de . On a : i

$$X\,G\,j_{\alpha\alpha}\lambda^{-} = \sum$$

$$^{1/2}(\,)\,(\,)\,F_{ij}$$

et

$$^{1/2}(\,)\,(\,)\,G\,j\,X_{ij}F\,i_{\alpha\alpha}\lambda^{-} = \sum$$

Il faut signaler que ces formules ne sont pas barycentriques comme celles du § 6.2.4 de l'analyse factorielle des correspondances ; les X_{ij} pouvant être négatifs.

5.2.5- Analyse des points supplémentaires

On profite de ce paragraphe pour parler éléments supplémentaires qui présentent un grand intérêt en analyse de données et plus particulièrement en analyse factorielle des correspondances. On utilise les éléments supplémentaires en analyse de représenter [ ] 14 :

- soit une observation relevée dans des conditions douteuses (ou différentes des autres données pour observations) ou encore une variable sur laquelle la précision est moindre que sur les autres variables mesurées ;
- soit un élément aberrant, ou ayant perturbé une analyse préliminaire ; 19
- soit un cas nouveau ;
- soit des éléments de nature différente de ceux analysés.

On peut aussi utiliser des éléments supplémentaires pour représenter un groupe de variables ou un groupe d'individus.

Exemple 1 : un questionnaire a été soumis à l'ensemble des étudiants du CASP ; après analyse, on recueille les réponses d'un étudiant absent (cas nouveau) : on cherchera ''naturellement'' à le placer sur les axes factoriels sans refaire l'analyse. Exemple 2 : on a réalisé une enquête sur l'image de marque de la S.N.E. Chaque client enquêté répond à un questionnaire comportant deux parties : une fiche socio démographique (âge, sexe, profession, revenus,…) ; et une batterie d'opinions relatives à

la
société. Si l'on analyse la batterie d'opinions, on mettra par exemple les variables
socio-démographiqu supplémentaires.
es en
C
onsidérons la figure suivante :

$J_s J$

$X_{ij}{}_s X_{ij}$

$_s X_{ij}$

$_s I$

Si l'on effectue l'analyse en composan incipales du tableau X_{ij} (tableau principal),
tes pr

on peut projeter sur les axes factoriels ainsi trouvés les ensembles $_s I$ (ensemble des individus supplémentaires) et $_s J$ (ensemble des variables supplémentaires). Les coordonnées des individus supplémentaires $_{s\,s} i\, I \in$ sont les composantes du vecteur $'(\,)_s$ $X u_{ij\alpha}$ et, celles des variables supplémentaires $_{s\,s} j\, J \in$ les composantes du vecteur $'(\,)_s X$ $v_{ij\alpha}$ (voir §4). ''Techniquement, mettre des éléments en supplémentaires dans l'analyse consiste à attribuer une masse nulle à ces éléments et à calculer leurs coordonnées dans l'espace factoriel''.

5.3. Interprétation de l'Analyse en Composantes Principales 5.3.1. Tableau des données de base

Traitons par cette méthode le recueil d'informations qui est donné par le tableau 1. Nous y trouvons les notes moyennes par matière obtenues par les étudiants du CASP promotion 1991-1992 pendant la première année de leur scolarité.

ABDO 10 15 13 14 12 17 10 10 10 12 09 09 07 13 12 08 14 13 12 BANZ 09 12 09 13 07 13 08 07 09 08 06 10 10 10 10 08 12 13 10 BATA 09 14 12 11 11 14 08 07 11 13 10 09 09 14 12 06 15 14 11 BOUK 11 15 11 13 10 18 11 08 13 11 13 15 11 13 14 07 13 12 12 BOYE 09 13 11 11 11 15 08 02 10

09 07 05 09 15 15 07 12 12 10 GOYI 09 13 12 13 11 15 10 11 14 14 07 08 12 14 15 07 14 14 12 LIK1 10 17 12 10 05 15 09 07 08 08 13 07 08 13 13 07 15 14 11 LIK2 12 14 15 13 11 18 08 11 12 10 08 11 10 13 10 07 15 15 12 LOUZ 06 14 07 13 09 13 11 14 11 07 07 09 06 12 10 08 13 14 11 MAKI 10 16 12 13 10 13 08 06 12 14 09 09 12 12 13 07 13 14 12 MALO 07 13 14 12 16 16 11 11 12 10 08 09 10 13 13 07 12 12 12 MAMP 10 14 13 11 13 13 13 13 10 12 10 07 10 14 13 08 13 13 12 MATO 10 14 16 12 06 15 10 13 13 09 09 09 09 13 11 08 15 14 12 MBIK 08 13 07 12 10 14 10 13 12 11 04 09 09 14 12 07 13 14 11 MPOU 09 15 10 13 09 15 09 08 11 13 08 07 11 13 17 07 14 14 12 NGUI 11 13 12 13 10 18 09 07 09 07 09 07 09 13 14 07 14 14 11 NKOK 09 17 11 13 09 14 11 07 11 11 12 09 11 14 13 08 15 14 12 NSEM 09 14 10 12 07 17 07 11 13 12 12 13 09 11 14 07 15 15 12 NSON 10 17 12 13 15 15 09 11 12 12 14 08 08 12 13 07 13 14 13 NZAK 09 16 10 13 14 15 07 05 09 11 17 11 12 14 13 08 14 13 12 ONDZ 10 17 12 10 05 15 09 07 08 08 13 07 08 14 13 07 15 14 11 SAFO 11 16 08 12 09 15 11 10 11 11 12 10 08 13 13 07 14 14 12 SAM1 12 15 10 14 15 17 07 07 12 08 09 09 08 10 11 08 14 14 12 SAM2 11 14 12 11 15 16 10 08 11 09 06 08 06 12 15 07 13 12 11 TSIB 10 15 17 13 11 15 06 06 12 10 12 09 11 13 15 07 15 14 12 Tableau.1 : Notes des étudiants

Le chef de la scolarité du CASP peut être amené à se demander :

- si les étudiants ont systématiquement des résultats meilleurs que ceux de leurs collègues ;
- si les filles et les garçons obtiennent des résultats comparables ;
- si un étudiant bon en mathématique l'est également en démographie ;
- etc.

Disons tout simplement qu'il veut analyser les données dont il dispose. Le tableau que nous allons étudier croise 25 étudiants (en lignes) et 19 matières (en colonnes) : le

nombre $_{ij}x$ se trouvant à la croisée de la ligne i et de la colonne j est la note obtenue par l'étudiant i à l'épreuve j .

Ensemble J (dans l'ordre des colonnes du tableau)

STAT Cours de statistique et organisation de la statistique
MSTA Méthodes statistiques
STAS Statistique de la santé
STAE Statistique de l'éducation et de l'emploi
MATH Mathématiques
PROB Probabilités
ECON Economie
DEMO Démographie
INFO Informatique
GEOE Géographie économie
COME Comptabilité d'entreprise

COMN Comptabilité nationale
TEXP Techniques d'expression
ANGL Anglais
HIST Histoire du Congo
SPOR Activités sportives
STAG Stage pratique
APPG Appréciation du Directeur Général du Casp
MOYG Moyenne générale

Ensemble *I* (des lignes)

Les étudiants sont repérés par leurs noms (sigle de 4 caractères). Dans l'analyse on mettra SPOR et MOYG en éléments supplémentaires, pour la simple raison que la note de sport n'intervient pas dans le système de pondération et que la moyenne générale n'est en fait qu'un résumé de toutes les notes. Il n'y a pas d'individus supplémentaires dans l'analyse.

5.3.2. Matrice de corrélations des variables

Tous les coefficients du tableau 2 ont été multipliés par 1000. C'est une matrice symétrique, seul le triangle inférieur est édité.

STAT MSTA STAS STAE MATH PROB ECON DEMO INFO GEOE COME COMN TEXP ANGL STAT 1000

MSTA 262 1000

STAS 341 43 1000

STAE 78 -80 -64 1000

MATH 68 -129 123 323 1000

PROB 544 -52 271 243 185 1000

ECON -204 -38 -139 -172 39 -180 1000

DEMO -254 -210 -54 84 1 -62 563 1000

INFO -12 -225 154 394 242 180 44 428 1000

GEOE -49 131 64 177 212 -159 60 109 472 1000

COME 228 751 153 -63 -60 136 -182 -264 -215 109 1000 COMN 112 9 -112 373 -30 354 -83 192 434 180 288

1000 TEXP -8 -8 208 297 7 -87 -164 -248 251 484 195 225 1000 ANGL -173 90 187 -397 -29 -114 297 -110 -146

246 126 -331 229 1000 HIST 46 108 112 -166 104 141 -40 -367 76 360 141 -245 305 346 STAT 325 442 336 -87

-426 236 -290 0 5 58 435 113 9 143 APPG 118 242 -16 139 -431 -18 -319 278 163 116 121 61 22 -164

HIST STAG APPG

HIST 1000

STAG -5 1000

APPG -232 693 1000

Tableau 2 : matrice de corrélation entre les variables

L'examen de cette matrice fait apparaître, avant toute analyse, les associations entre variables. Nous remarquons que le coefficient le plus élevé vaut 0.751, c'est celui de la comptabilité d'entreprise (COME) avec les méthodes statistiques (MSTA) ; ensuite vient celui de stage pratique (STAG) avec l'appréciation générale (APPG) qui est égal à 0.693. On peut signaler des coefficients de corrélation moyens entre la démographie (DEMO) et l'économie (ECON) soit 0.563 et entre statistique (STAT) et probabilités (PROB) soit 0.544. De façon générale on constate entre les variables, des coefficients de corrélation faibles : un coefficient de corrélation nul entre la démographie (DEMO) et le stage pratique (STAG) et un coefficient quasi nul entre les mathématiques (MATH) et la démographie (DEMO). En d'autres termes connaissant la valeur DEMO, on ne peut rien dire des valeurs STAG et MATH. Dans le nuage où les points sont des variables et les droites des individus, DEMO fait un angle droit avec STAG et avec MATH. Il faut néanmoins rappeler qu'un coefficient de corrélation est un indice qu'il faut interpréter avec beaucoup de précautions.

C'est donc en définitive cette matrice de corrélation entre variables qu'il faut diagonaliser pour obtenir une représentation euclidienne des points variables.

5.3.3. Vecteurs et valeurs propres de la matrice de corrélation

Les vecteurs

propres contiennent les informations à affecter aux variables initiales et qui permettent le calcul des facteurs. Le tableau 3 rend donc compte de ce passage de l'ancien repère R_J des anciennes variables au nouveau repère R_F des nouvelles variables (facteurs).

Comme on l'a déjà noté la somme des valeurs propres est égale au nombre $Card_J$ de variables. Dans le cas d'un nuage sans direction d'allongement (nuage sphérique), toutes les valeurs propres seraient égales à 1. Ce cas limite permet de retenir comme axe à priori à étudier ceux dont les valeurs propres sont supérieures à l'unité (les six premières dans le cas présent). La valeur 1 constitue donc un point de repère pour apprécier une valeur propre. Dans l'interprétation d'un facteur associé à une valeur propre proche ou inférieure à 1 il est conseillé d'être très prudent. Ce qui vient d'être dit n'est valable que si on travaille sur les données centrées et réduites (ACP normée c'est-à-dire la méthode jusque ici développée).

NUMERO ! VAL PROPRE 1! VAL PROPRE 2 ! VAL PROPRE 3 ! VAL PROPRE !

! -. 3.07452 ! 2.66202 ! 2.32011 ! 1.96808 !

OBJET 1 ! - .33048 ! .06183 ! -0.3186 ! .32050 ! OBJET 2 ! -.36071 ! -.19292 ! -.02134 ! -13812 ! OBJET 3 ! -22630 ! 01539 ! 17381 ! 08163 ! OBJET 4 ! -04945 ! 45327 ! 02122 ! 10787 ! OBJET 5 ! 12295 ! 20650 ! 31472 ! 27844 ! OBJET 6 ! -22630 ! 21435 ! -00401 ! 37095 ! OBJET 7 ! 29212 ! -04166 ! 05790 ! -26369 ! OBJET 8 ! 22630 ! 21435 ! 00401 ! 37095 ! OBJET 9 ! 00108 ! 48566 ! 14409 ! -19273 ! OBJET 10 ! -09790 ! 20351 ! 38798 ! -36938 ! OBJET 11 ! -41111 ! -12271 ! 05832 ! -05264 ! OBJET 12 ! -14995 ! 40369 ! -07915 ! -02468 ! OBJET 13 ! -16463 ! 14701 ! 39255 ! -16085 ! OBJET 14 ! -01500 ! -29097 ! 33802 ! -26204 ! OBJET 15 ! -10733 ! -13306 ! 46651 ! 01130 ! OBJET 16 ! -44550 ! -05062 ! -19581 ! -22215 ! OBJET 17 ! 27514 ! 10600 ! -33571 ! -33841 ! Tableau 3 : Les quatre premiers vecteurs et valeurs propres.

On donne ci-dessous l'histogramme des valeurs propres qui permet de visualiser l'importance et la décroissance des valeurs propres. On note une décroissance lente des valeurs propres.

On peut faire remarquer que :

$$_{121}(\lambda - = \lambda \lambda)/ (3.07452 - 2.66202)/ 3.07452 = 0.13417$$

$$_{232}(\lambda - = \lambda \lambda)/ (2.66202 - 2.32011)/ 2.66202 = 0.12844$$

$_{3\,4\,3}(\lambda - = \lambda\ \lambda\ )/\ (2.32011 - 1.96808)/\ 2.32011 = 0.15173$

Si l'écart relatif entre λ_α et $\lambda_{\alpha+1}$ est faible, une légère fluctuation dans le tableau des données peut avoir pour conséquence la permutation des facteurs d'ordre α et α +1. En règle générale ''si des valeurs propres successives sont proches l'une de l'autre, on considérera le sous-espace défini par les axes associés à ces valeurs propres, et non les axes séparément. En effet, il s'agit dans un tel cas pratiquement d'un sous-espace propre, et la position des axes dans ce sous-espace n'est pas significative : elle est définie à une rotation près''[60].

LES VALEURS PROPRES VAL (1) = 3.07452

NUM	VAL PROPRE	POURC.	CUMUL	VARIAT.	*HISTOGRAMME DES VALEURS PROPRES
1	3.07452	18.085	18.085	*******	*
2	2.07452	15.659	33.744	2.426	*
3	2.32011	13.648	47.392	2.011	*
4	1.96808	11.577	58.969	2.071	*
5	1.48126	8.713	67.682	2.864	*
6	1.31772	7.751	75.434	.962	*
7	.90442	5.320	80.754	2.431	*
8	.80785	4.752	85.506	.568	*
9	.59308	3.489	88.994	1.263	*
10	.56950	3.350	92.344	.139	*
11	.44687	2.629	94.973	.721	*
12	.27241	1.602	96.576	1.026	*
13	.23710	1.395	97.970	.208	*
14	.14339	.843	98.814	.551	*
15	.10191	.599	99.413	.244	*
16	.06428	.378	99.791	.221	*
17	.03548	.209	100.000	.169	*

Tableau 4 : Histogramme des valeurs propres

Les deux premières valeurs propres représentent environ 34% de l'inertie et les six premières environ 75%. Notons que ces taux sont faibles. Du fait des faibles coefficients de corrélation entre les variables on ne pouvait pas s'attendre à trouver des valeurs propres très élevées.

Il faut avouer qu'il est difficile de donner une réponse générale à la question : à partir de quel pourcentage d'inertie peut-on négliger les facteurs restants ? Cela dépend en général

du nombre de variables : un % de 100% n'a pas le même intérêt sur un tableau de 20 variables et sur un tableau de 100 variables[57]. Cependant des taux d'inertie faibles peuvent aussi donner des représentations de bonne qualité. On s'assurera néanmoins qu'un fort pourcentage d'inertie est presque une garantie ''d'interprétabilité au premier sens du terme''. Nous essayerons de résumer les données par les trois premiers facteurs.

5.3.4. Tableau des facteurs sur I

Le tableau 5 est en fait un tableau d'aide à l'interprétation d'une analyse en composantes principales (comme d'ailleurs les tableaux 6 et 7).

! I ! QLT POID INR ! 1#F COR CTR ! 2#F COR CTR ! 3#F COR CTR !

1 ! ABDO! 279 40 20 ! 48 0 0 ! 763 67 9 ! -140 2 0 ! 2 ! BANZ! 799 40 63 ! 2563 245 86 ! 400 6 2 ! -1998 148 69 ! 3 ! BATA! 233 40 20 ! -450 24 3 ! -830 81 10 ! 50 0 0 ! 4 ! BOUK! 383 40 55 ! -856 31 10 ! 2158 198 70 ! 1630 113 46 ! 5 ! BOYE! 839 40 61 ! 1964 148 50 ! -2899 322 126 ! 2308 204 92 ! 6 ! GOYI! 786 40 37 ! 512 17 3 ! 1711 186 44 ! 2248 321 87 ! 7 ! LIK1! 955 40 48 ! -1693 140 37 ! -3696 665 205 ! -1726 145 51 ! 8 ! LIK2! 745 40 45 ! -1972 202 51 ! 2278 270 78 ! -1435 107 36 ! 9 ! LOUZ! 885 40 70 ! 3683 453 177 ! 213 2 1 ! -3210 344 178 !
10 ! MAKI! 613 40 29 ! -708 40 7 ! 781 49 9 ! 1290 133 29 ! 11 ! MALO! 739 40 45 ! 2923 442 111 ! 883 40 12 ! 1899 187 62 ! 12 ! MAMP! 670 40 39 ! 1901 216 47 ! -1104 73 18 ! 1343 108 31 ! 13 ! MATO! 631 40 32 ! -375 10 2 ! 624 29 6 ! -1488 165 38 ! 14 ! MBIK! 786 40 37 ! 2920 542 111 ! 620 24 6 ! -533 18 5 ! 15 ! MPOU! 432 40 26 ! -309 9 1 ! -233 5 1 ! 1685 253 49 ! 16 ! NGUI! 507 40 30 ! -429 15 2 ! -664 35 7 ! -600 29 6 ! 17 ! NKOK! 599 40 24 ! -1275 163 21 ! -720 52 8 ! 492 24 4 ! 18 ! NSEM! 576 40 46 ! -1848 174 44 ! 2132 231 68 ! -1418 102 35 ! 19 ! NSON! 96 40 27 ! -658 38 6 ! 749 50 8 ! 288 7 1 ! 20 ! NZAK! 630 40 46 ! -1915 187 48 ! -608 19 6 ! 1614 133 45 ! 21 ! ONDZ! 976 40 50 ! -1706 136 38 ! -3936 724 233 ! -1446 98 36 ! 22 ! SAFO! 230 40 18 ! -458 28 3 ! -283 11 1 ! -822 89 12 ! 23 ! SAM1! 813 40 50 ! -913 39 11 ! 2075 202 65 ! -1825 156 57 ! 24 ! SAM2! 736 40 42 ! 1732 170 39 ! -655 24 6 ! 587 19 6 ! 25 ! TSIB! 575 40 37 ! -2683 460 94 ! 242 4 1 ! 1204 93 25 ! ! ! 1000 ! 1000 ! 1000 ! 1000 !

Tableau 5 : Facteurs sur I

Pour chacun des 25 étudiants on lit d'abord :

i) $POID$ (masse statistique) : on constate que tous les individus ont reçu le même poids.

$$m$$

$$1 \sum = ; m_i = ; = 25\ M\ m_i^{1}\ 0{,}04$$

p_M= millièmes).

$\frac{i}{i}$ exprimé en 25

=–= (ici 40

ii) R (inertie) ; les individus ayant le même poids, cette inertie varie comme la distance IN au centre de gravité :

$^2( ) ( )_{n\,i} I\, i\, p$ = ρ i

27

iii) LT , sa qualité de représentation par sa projection dans l'espace factoriel considéré Q comme significatif.

Ensuite on trouve pour chaque facteur :

$\#\,(F\,F\,(i))_{\alpha}$

iv) , coordonnées des individus ; l'examen de ces coordonnées permet de connaître comment se répartissent les individus, ceux qui interviennent sur l'axe du côté positif ou du côté négatif.

v) CTR , contribution relative de l'individu i et à l'inertie expliquée par l'axe α . $^2($, $) ( )/ CTR_{\,i}\, i\, p\, F\, i$ α = $_{\alpha}$λ$_{\alpha}$;

on remarque que CTR varie comme i) : les points les plus contributifs sont les plus $^2($

F_{α}

excentrés et réciproquement. La contribution relative de l'étudiant LOUZ à l'inertie expliquée par l'axe 1 est égale à 177. En d'autres termes si on appelle 1000 le facteur 1, LOUZ en explique 177. Pour l'interprétation des axes, on classera les individus en deux groupes ; les uns de contribution relative forte avec une coordonnée négative, les autres de contribution forte avec une coordonnée positive (il est conseillé de choisir les individus de contribution relative supérieure à la moyenne des contributions au moins). *vi)* COR qui mesure la qualité de la représentation de l'individu i par sa projection sur l'axe cosinus de l'angle formé par un point avec sa α . COR peut être interprétée comme le projection sur le plan.

ρ( )i

Gθ

^{2}
s θ $_2$( )
i

$F\,i$( ) $_{\alpha}$

$F\ i$

$\rho =$

^{2}

co( )

α

i

En prenant toujours le cas de l'étudiant LOUZ on voit que $^{2}{}_{1} \cos (LOUZ, F) = 453$; si on appelle 1000 la situation de LOUZ, on en trouve 453 sur le facteur 1. On peut vérifier $_{p}$ facilement que : ( ) 1000

= = $\sum$ (en millième). Comme on le constate

les COR $QLT\ COR\ i_{\alpha}$

$\alpha =$ 1

s'additionnent en lign n e ; sommés sur les 17 facteurs, o trouverait 1000 ; sommés sur les 5 (on a extrait 5 facteurs, seulement trois sont imprimés) facteurs, on trouve pour LOUZ..

$QLT = 885\ 28$

5.3.5 Tableaux de facteurs sur J .

! J1 ! QLT POID INR ! 1#F COR CTR ! 2#F COR CTR ! 3#F COR CTR !

1 ! STAT ! 661 1 59! -579 336 109 ! 101 10 4 ! -49 2 1 ! 2 ! MSTA ! 546 1 59 ! -632 400 130 ! -315 99 37 ! -33 1 0 ! 3 ! STAS ! 540 1 59 ! -397 157 51 ! 25 1 0 ! 265 70 30 ! 4 ! STAE ! 665 1 59 ! -87 8 2 ! 740 547 205 ! 32 1 0 ! 5 ! MATH! 560 1 59 ! 216 46 15 ! 337 114 43 ! 479 230 99 ! 6 ! PROB ! 746 1 59 ! -397 157 51 ! 350 122 46 ! -6 0 0 ! 7 ! ECON ! 621 1 59 ! 512 262 85 ! -68 5 2 ! 88 8 3 ! 8 ! DEMO! 894 1 59 ! 356 127 41 ! 431 186 70 ! -332 110 47 ! 9 ! INFO ! 784 1 59 ! 2 0 0 ! 792 628 236 ! 219 48 21 ! 10! GEOE ! 765 1 59 ! -172 29 10 ! 332 110 41 ! 591 349 151 ! 11! COME ! 590 1 59 ! -721 520 169 ! -200 40 15 ! 89 8 3 ! 12! COMN! 541 1 59 ! -263 69 22 ! 659 434 163 ! -121 15 6 ! 13! TEXP ! 676 1 59 ! -289 83 27 ! 240 58 22 ! 598 358 154 ! 14! ANGL ! 716 1 59 ! -26 1 0 ! -475 225 85 ! 515 265 114 ! 15! HIST ! 589 1 59 ! -188 35 12 ! -217 47 18 ! 711 505 218 ! 16! STAG ! 859 1 59 ! -781 610 198 ! -83 7 3 ! -298 89 38 ! 17! APPG ! 753 1 59 ! -482 233 76 ! 173 30 11 ! -511 261 113 ! ! ! 1000 ! 1000 ! 1000 ! 1000 !

Tableau 6 : Facteurs sur J .

On a donné à chaque point variable une masse égale à l'unité et que les coordonnées factorielles de ces points sont assimilables aux coefficients de corrélation. On a

$^{2} COR(\ , j)\ G\ (j)$ α = α variables qui lui sont

. On interprètera donc l'axe en fonction des

corrélées. Comme le nuage $(\ )_{I} j$ est situé dans une sphère de rayon 1, l'usage des CTR

N

n'est pas vraiment nécessaire. On retiendra seulement que plus la variable se projette près du cercle dans le plan principal, mieux cette variable est représentée par sa projectio

n

.

! JSUP ! QLT POID INR ! 1#F COR CTR ! 2#F COR CTR ! 3#F COR CTR ! 18 ! SPOR ! 141 1 59 ! 179 32 0 !

114 13 0 ! -260 68 0 ! 19 ! MOYG ! 669 1 59 ! -454 206 0 ! 520 270 0 ! 248 62 0 ! ! 118 ! 0 ! 0 ! 0 !

Tableau 7 : Facteurs sur J supplémentaires.

5.3.6. Représentations graphiques

Le but essentiel de l'analyse factorielle est de représenter les points de et de $(\ )\ N_{J} J$

$(\ )\ N_{J} I$

dans un espace de faible dimension par rapport aux dimensions d'origine. Ces représentations se font dans la plupart **aux points** i

des cas dans un espace à d **5.3.6.1.** eux dimensions :

Représentation graphique associée

Les graphiques 1 et 2 donnent une représentation des individus dans l'espace factoriel (1,2) et (1,3). Si l'on s'est fixé comme objectif la répartition des individus, on peut interpréter rapidement les résultats de la façon suivante : l'axe 1 oppose les individus BANZ, MBIK et LOUZ à l'individu TSIB (voir leurs coordonnées, tableau 5). Cet axe oppose en fait les étudiants ‘’reçus’’ et non ‘’reçus’’. Une exception : l'étudiant MALO qui est reçu

mais qui se retrouve avec les non reçus. On retrouve la même répartition sur l'axe 2 : opposition entre BOYE, LIK1, ONDZ (groupe des non reçus) et GOYI, BOUK, LIK2 et SAM1 (groupes des reçus). On peut donc considérer ces axes comme des axes de ''réussite''.

Ensuite, non loin de l'origine des axes, on constate des groupements d'individus selon toujours le critère ''réussite'' : NZAK avec NKOK (reçus), NGUI avec BATA (non reçus), MAKI, ABDO et MATO (reçus). On peut le vérifier aussi pour les points superposés indiqués en bas du graphique.

Cependant compte tenu des remarques faites au §5.3.3 on est tenté d'analyser le plan (1,3). A quelques exceptions près on retrouve la même interprétation.

Notons que dans cette analyse des individus, l'origine des axes représente '' l'individu moyen'', dont les notes sont les moyennes calculées sur l'ensemble des étudiants.

Le cas que nous venons d'examiner est celui où les individus présentent de l'intérêt en eux mêmes. L'information essentielle est contenue dans les coordonnées. Dans d'autres cas, en particulier lorsque les individus constituent un échantillon (situation typique des enquêtes) on est en présence des êtres anonymes n'ayant d'intérêt que par leur ensemble et non par leur individualité 0]. L'attention sera alors attirée par l'allure générale de la répartition

[1

de l'ensemble des individu

s

.

5.3.6.2. Représentation graphique associée aux points *j*

Les graphiques 3 et 4 donnent une représentation des points variables dans les plans factoriel (1,2) et (1,3). On peut se fixer comme l'objectif la structuration des variables : quelles sont celles qui sont associées ? Quelles sont celles qui s'opposent ?

Un simple regard de leurs coordonnées sur le premier axe nous indique que la plupart des variables sont d'un même côté (côté négatif). Deux variables sont bien corrélées avec le premier facteur : il s'agit de STAG et COME. Des variables moyennement corrélées avec le premier facteur : STAT, MSTA et APG. Du côté positif de l'axe on peut retenir la variable ECON moyennement corrélée avec le premier facteur. De façon générale, l'axe 1

peut donc être considéré comme axe de la ‘’pratique’’.

Le deuxième facteur est corrélé positivement avec l’informatique (INFO), les statistiques de l’éducation (STAE) et la Comptabilité nationale (COMN). Du côté négatif de l’axe, se trouve la variable anglais (ANGL).

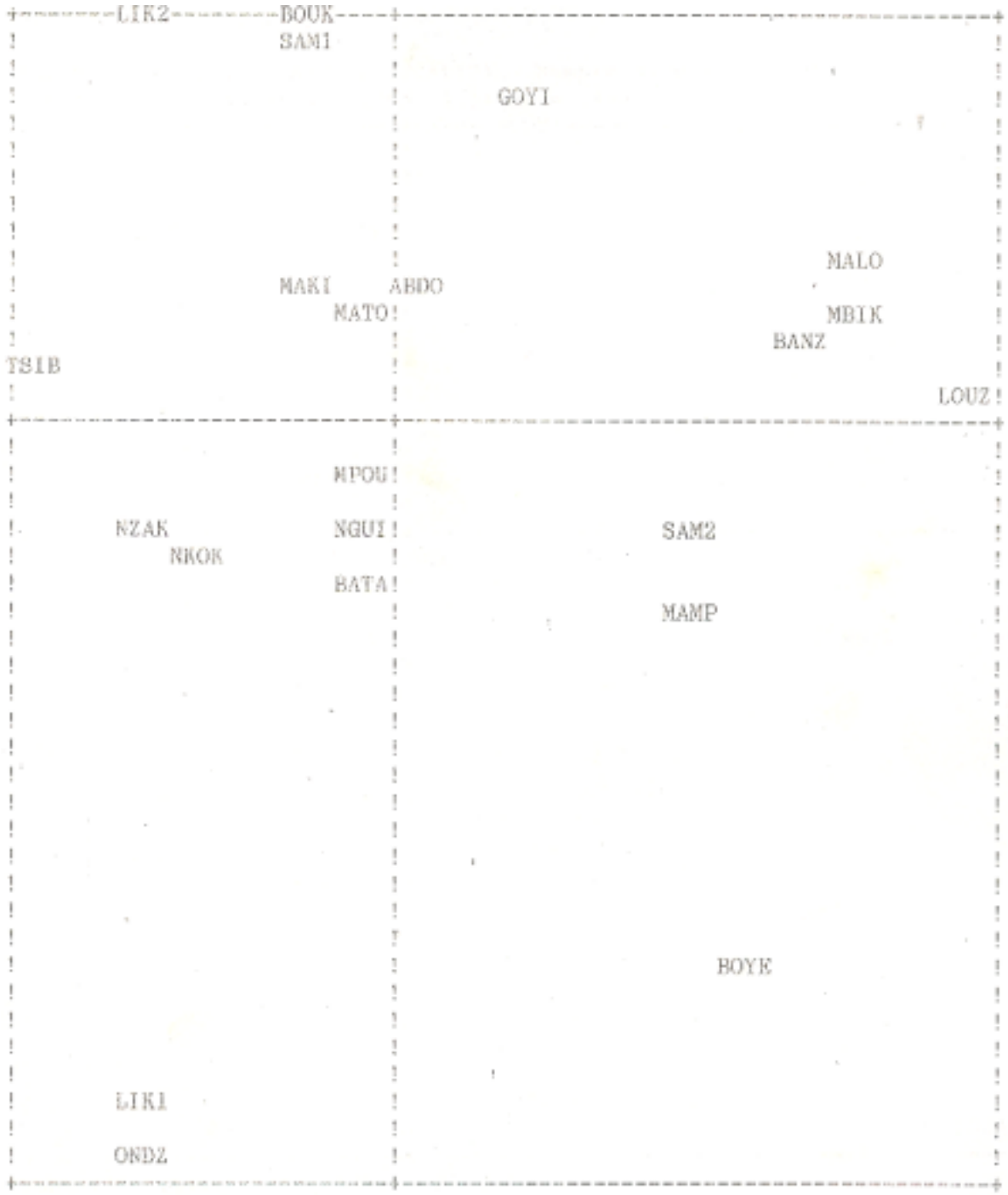
Le troisième facteur peut être considéré comme facteur de culture générale. Sont effectivement corrélées avec ce facteur les variables histoire (HIST), anglais (ANGL), techniques d’expression (TEXP) et la géographie (GEOE).

5.3.6.3. Représentation simultanée des points i et des points j .

Bien que des individus et variables soient des éléments d’espaces différents, on peut par un certain artifice superposer la représentation des individus (plan principal) et celles des variables (cercle de corrélation). Une telle superposition ‘’avec des précautions d’interprétation, rend plus vivante la visualisation’’. Le graphique 5 est donc issu des graphiques 1 et 3. ‘’Ainsi, si l’on regarde simultanément les deux graphiques, un individu sera du côté des variables pour lesquelles il a de fortes valeurs et à l’opposé des variables pour lesquelles il a de faibles valeurs’’.

AXE HORIZONTAL (1) – AXE VERTICAL (2). Nombre de points : 25

Echelle : 4 caractères = .354 1 ligne = .147

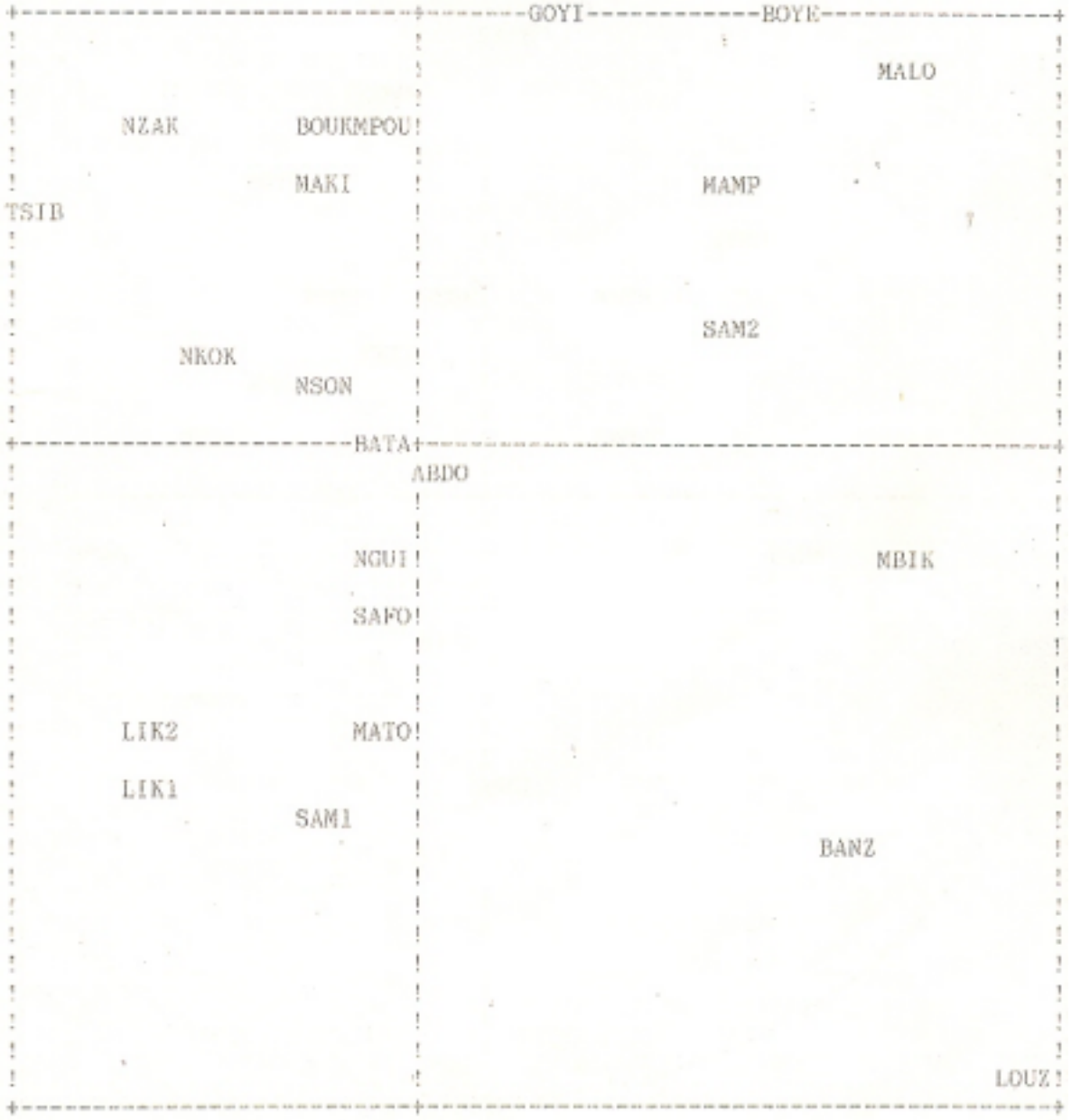


Nombre de points superposés : 3
NSEM(LIK2) NSON(MAKI) SAFO(MPOU)

Graphique 1 : Représentation des points individus dans l'espace factoriel (1,2). 32

AXE HORIZONTAL (1) – AXE VERTICAL (3). Nombre de points : 25

Echelle : 4 caractères = .354 1 ligne = .147



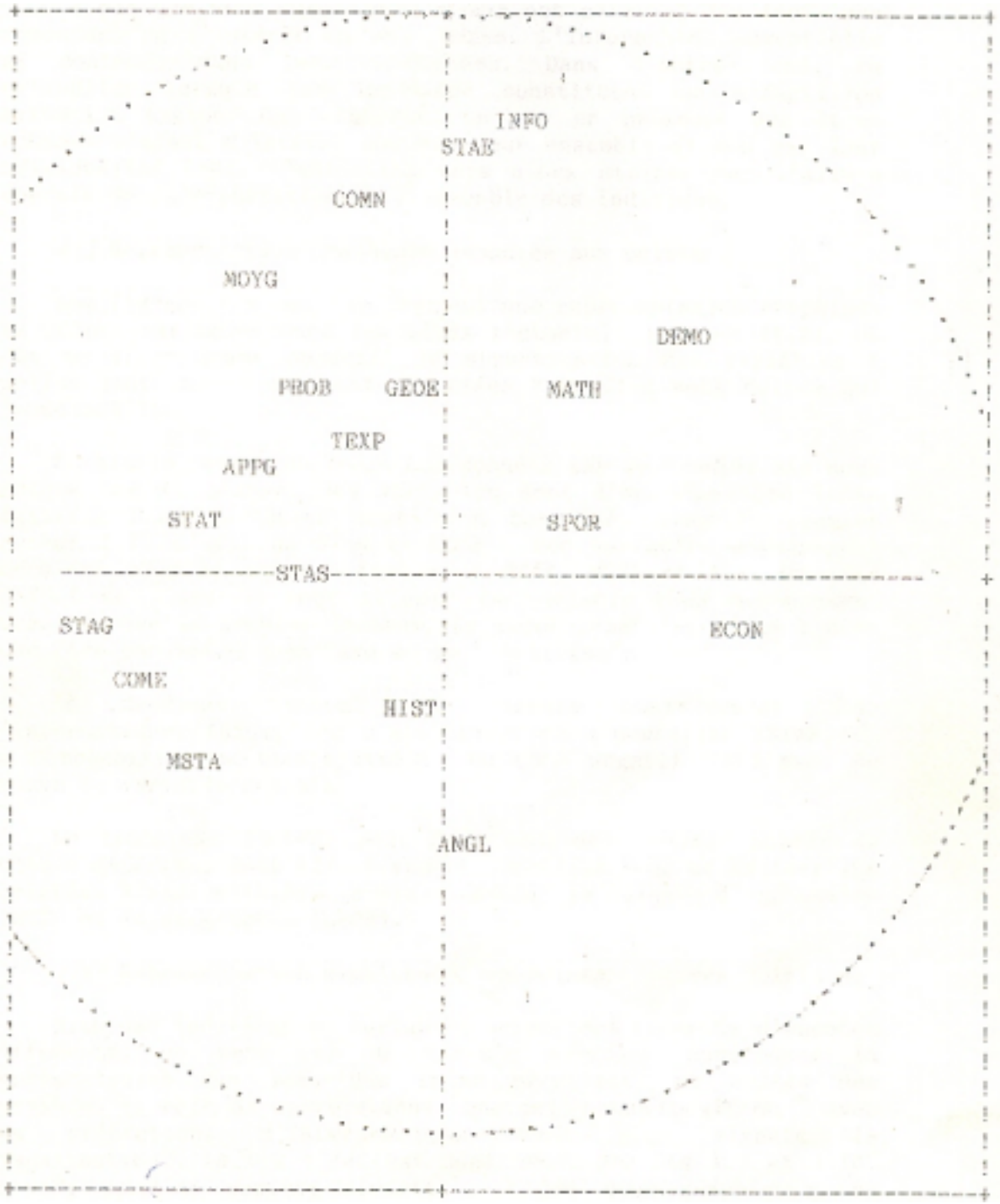
Nombre de points superposés : 2

NSEM(LIK2) ONDZ(LIK2)

Graphique 2 : Représentation des points individus dans l'espace factoriel (1,3). 33

AXE HORIZONTAL (1) – AXE VERTICAL (2). Nombre de points : 31

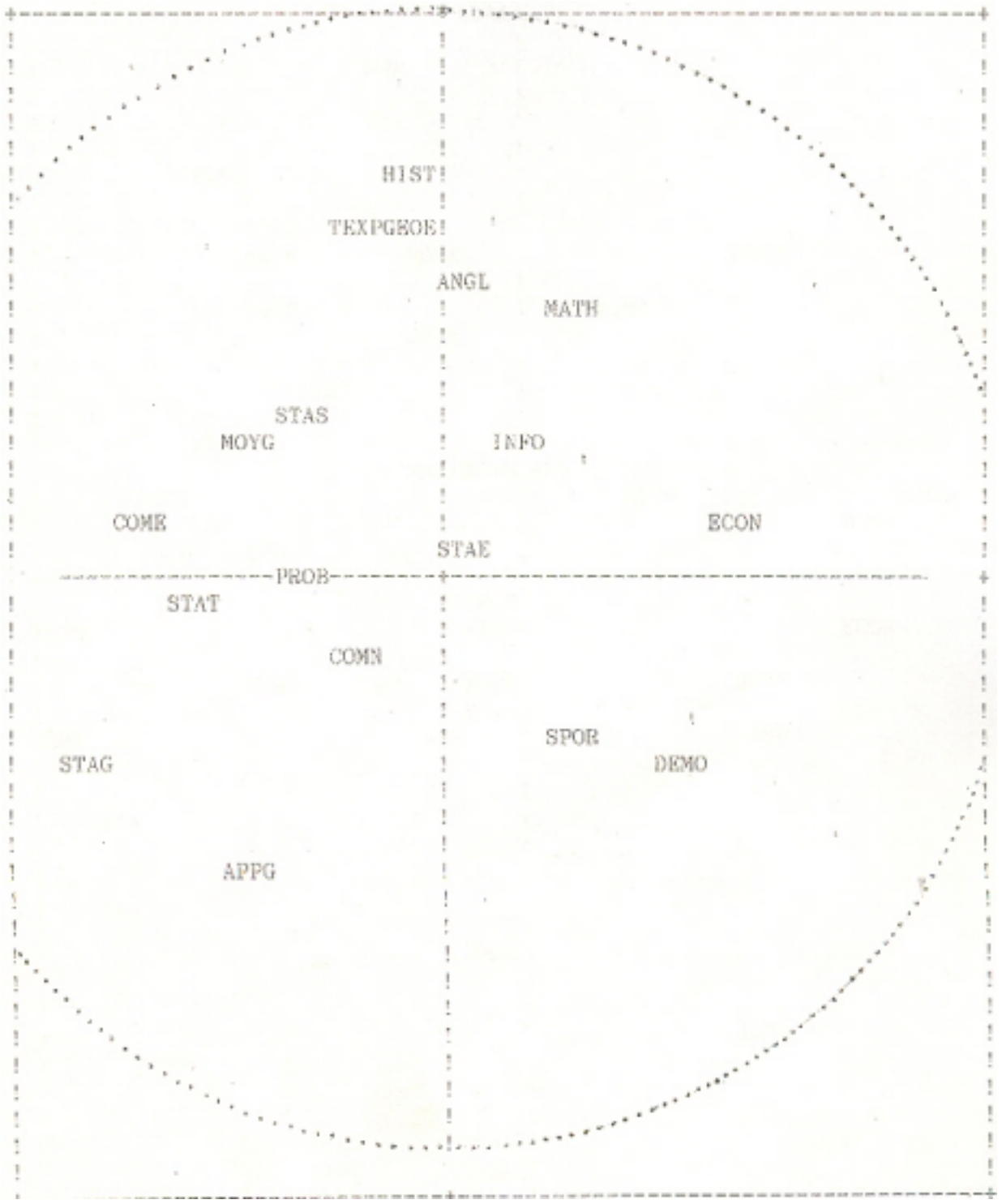
Echelle : 4 caractères = .111 1 ligne = .046



Graphique 3 : Représentation des points variables dans l'espace factoriel (1,2)

AXE HORIZONTAL (1) – AXE VERTICAL (3). Nombre de points : 31

Echelle : 4 caractères = .111 1 ligne = .046

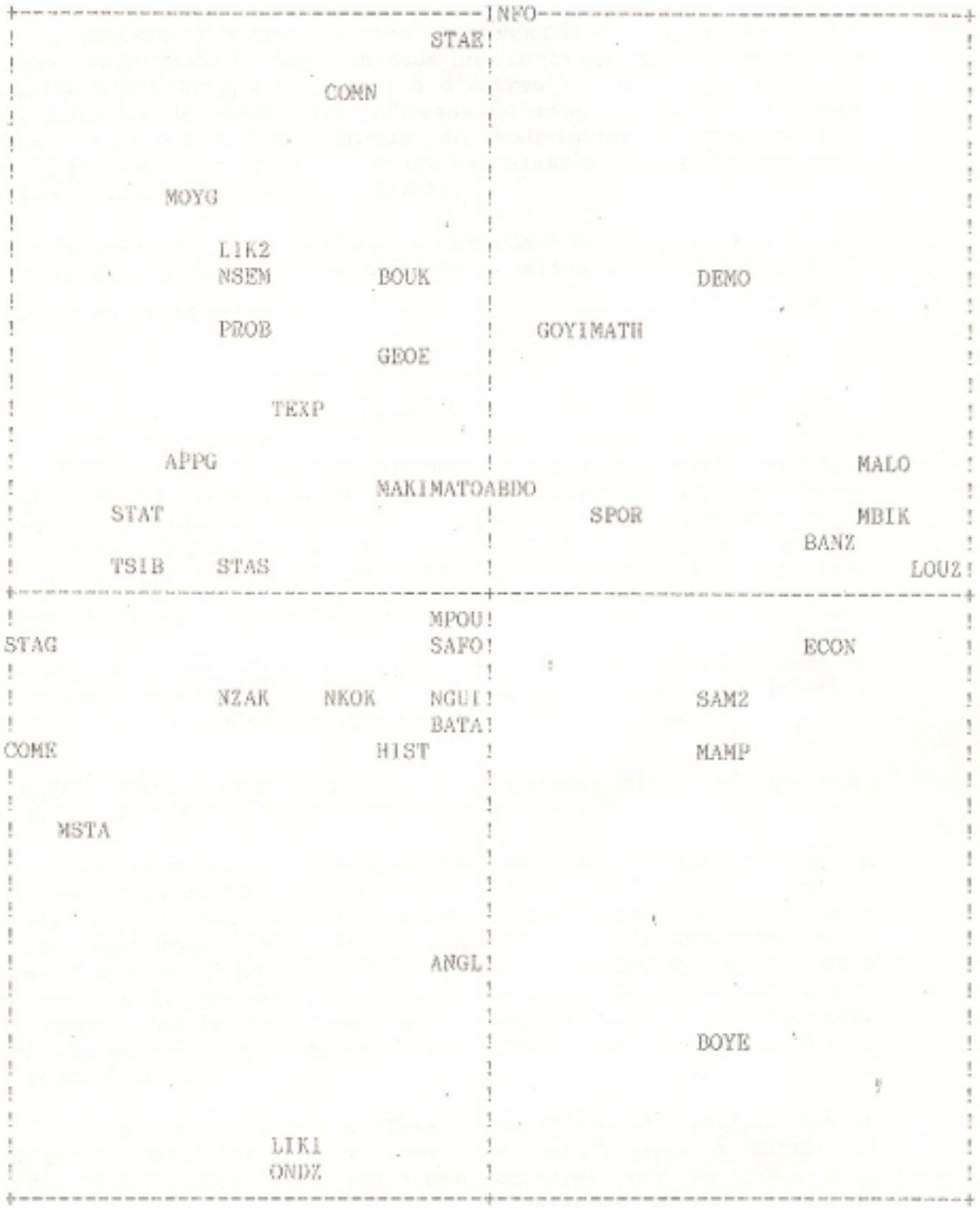


Nombre de points superposes: 1
MSTA (STAT)

Graphique 4:. Représentation des points variables dans l'espace factoriel (1,3) 35

AXE HORIZONTAL (1) – AXE VERTICAL (2). Nombre de points : 44

Echelle : 4 caractères = .422 1 ligne = .176



Nombre de points superposés : 2

SAM1 (BOUK) NSON(MAKI)

Graphique 5: Représentation simultanée des individus et des variables dans l'espace factoriel (1,2).

6. L'analyse factorielle des correspondances

6.1. Les données – Les objectifs

L'analyse factorielle des correspondances a d'abord été conçue pour traiter les tableaux de contingence ; depuis, son domaine s'est très vite étendu à d'autres tableaux de données : les tableaux de notes, les tableaux de rang… etc. Et récemment elle s'applique à des tableaux de description logique remplis exclusivement de 1 et de 0 ; c'est par exemple le cas des tableaux mis sous forme disjonctive complète.

En effet, si l'on considère un ensemble Q , de questions (ou de variables qualitatives), pour toute question de , on note l'union disjointe des $_q q\ Q\ J\ J$:

$\{\}\ _q J\ J$ = ∪ $q \in Q$ (avec $CardJ = p$).

Soit I ($ardI = n$) un ensemble d'individus ayant répondu à toutes les questions de Q . C

Pour tout i et de que l'individu i a adopté I , et pour toute question q de Q , on suppose une seule modalité de réponse à q , et l'on code par 1 si l'individu i a choisi la modalité de pons J de $_q$ré e J , et par 0 sinon. Le tableau ainsi obtenu est appelé disjonctif complet :

Disjonctif, car deux modalités j et $'j$ d'une même question s'excluent mutuellement : si l'individu i a choisi la modalité j de $_q J$, il n'a pas adopté une modalité $''j$ ($j \neq j$) de $_q J$.

Complet, car à tout individu correspond effectivement une modalité de réponse à toute i question q.

Contrairement à l'analyse en composantes principales, en analyse factorielle des correspondances (AFC) le tableau à analyser est symétrique par rapport aux indices i et j . Deux lignes sont considérées comme proches si elles ‘’s'associent de la même façon à l'ensemble des colonnes’’. Symétriquement, ‘’deux colonnes sont proches si elles s'associent de la même façon à l'ensemble des lignes’’. L'AFC permet donc de traiter simultanément les ensembles I et J et de les confronter en vue de découvrir l'ordre général. Enfin, comme l'ACP, l'analyse factorielle des correspondances permet de réaliser un (ou plusieurs) graphiques, à partir du tableau de données, ‘’en réduisant les dimensions de l'espace de représentation’’ des données, tout en essayant de ne pas perdre trop d'information au moment de cette réduction.

6.2. La méthode

Nous ne dirons rien sur la méthode ; nous bornant seulement à citer J.P Benzécri : ‘’ainsi une méthode unique dont le formulaire reste simple est parvenue à incorporer des idées et des problèmes nombreux apparus d'abord séparément, depuis plusieurs décennies…’’[6].

6.2.1. Le tableau des données

Soient deux ensembles finis I et J en correspondance : on a :

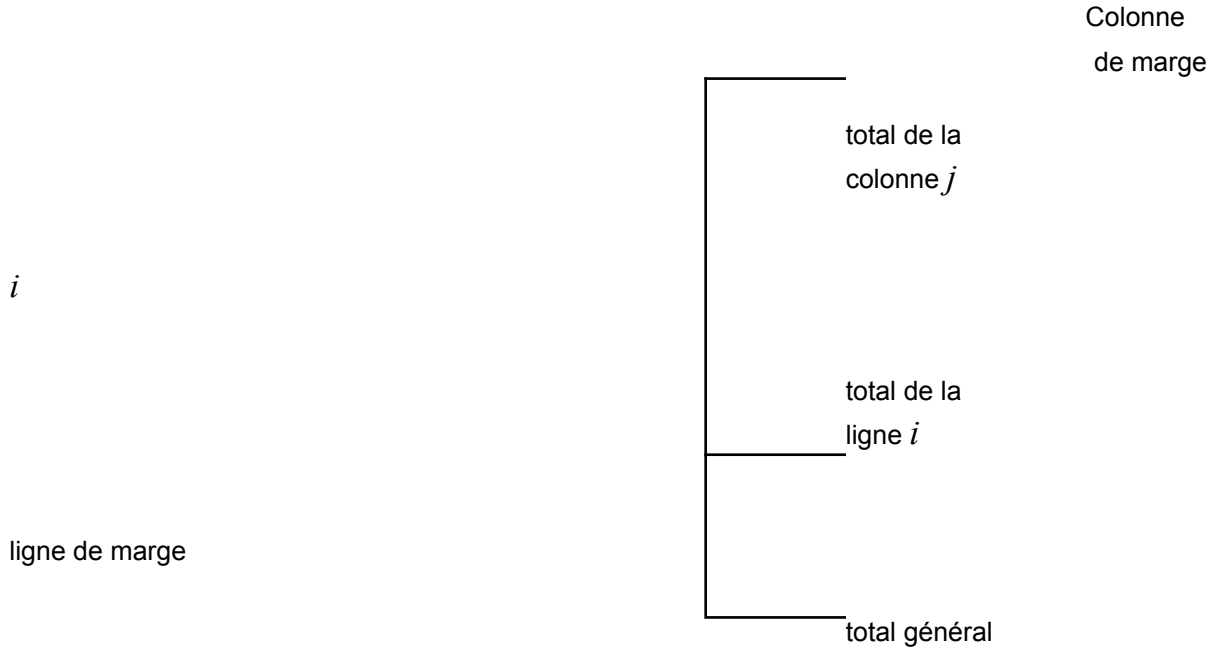
$\{ , \}_{IJ\,ij} k = \in k\ i\ I\ j \in J$, un tableau homogène de nombres sur le produit de ces deux ensembles I et J ($CardI = n$,$CardJ = p$). On pose :

$_{ij}k$ effectif de la case (i , j) ; $k \geq 0\ _{ij}$;

$\{ _{i\,ij} k\ k = j \in J\}$ l'effectif de la ligne i ; la colonne des éléments $_i k$ est la *colonne marginale ;*

$\{ \}_{j\,ij} k\ k = i \in I\ j$; la ligne des éléments $_j$ est l'effectif de la colonne k est la *ligne marginale ;*

$\{$ [redacted] $, \}_{i\,j\,ij}$ $k =$ [redacted] k [redacted] $\sum \sum$ [redacted] $i\ I \in j \in J$ est la somme du tableau. On a le schéma suivant :



D k ivisons maintenant chaque valeur du tableau précédent par (cardinal de la population).

$/_{ij\,ij} f = k\ k$: fréquence d'un couple $(i\ j\ ,\ )$.

$\{ \}_{i\,ij\,i} f = \in \sum f\ j\ J = k\ /\ k$ est la fréquence d'une ligne i ; la colonne des $_i f$ est la *colonne des fréquences marginales*.

$\{ \} /_{j\,ij\,j} f = \in \sum f\ i\ I = k\ k$ est la fréquence de la colonne j ; la ligne des $_j f$ est la *ligne des fréquences marginales.*

Par construction on a évidemment :

$\{$ [redacted] $\} 1 \sum_{ij} \sum f$ [redacted] $j \in J =$.

Définissons maintenant le *profil* d'un élément i de I et d'un élément j de J le tableau à $_{i}$ $_{j}f$ et $^{j}{}_{i}$ une dimension, noté respectivement f et dont le contenu est déterminé de la façon suivante :

$_{Jj}f = = fj, CardJ\}$, avec $_{i}/ / {}^{i}{}_{j\,ij\,ij}f = = ff\,k\,k$ et $0\ _{i}f \neq$; $0\ _{i}k \neq$; $^{i}{}_{j}f$ est la *fréquence*

$\{ 1,... {}^{i\,i}$

conditionnelle du couple $(,ij)$ connaissant i .

$\{ {}^{1,...,} \}_{jj}$

$_{I\,i}f = = fi\ Card$ avec $/ / {}^{j}{}_{i\,ij\,j\,ij}If = = ff\,k\,k$ et $0\ _{j}f \neq$; $0\ _{i}k \neq$; $^{i}{}_{j}f$ est la *fréquence conditionnelle* du couple $(,ij)$ connaissant j

$_{J}$Le tableau ^{i}f correspond au tableau des pourcentages en lignes. C'est donc le profil de la ligne i ; le tableau $^{j}{}_{i}f$ correspond au tableau des pourcentages en colonnes ; on parle alors de profil de la colonne j . On a :

$\forall \in i\ I : 1\ {}^{i}{}_{j} \sum f = ; \forall \in j\ J : 1\ {}_{j}{}_{i} \sum f =$

6.2.2. Analyse des points de dans

$$( ) N I\ {}^{p}{}_{J}\ i\ R$$

Dans l'espace des colonnes, le point i sera muni de la masse $_{i}f$ et représenté par son profil $_{J}f$ (sa composante sur la j è − me variable est $/ {}^{i}{}_{j\,ij\,i}f = ff$) ; on notera par :

$_{i}$

$( ) \{(,) \}\ {}^{i}N\ I_{J\,j\,i} = ff\,i \in I$, le nuage des points $i\ I \in$

$\{ ; \}_{Jj}$ ***i) Le centre de gravité, de ce nuage est*** $f = fj \in J$

r G d'un système, de points $\{(,) \}\ m\ x_{i\,ij}$ Comme le centre de g avité (ou barycentre) $i \in I$ est nt moyen du système, de la j è − me composante tel que pour tout $_{Gj}$ le poi Xj : $\{ ( ) \}\ 0$

$m\ x_{i\,ij\,Gj} \sum - \in x\ i\ I =$

On a donc pour tout i :

$$f_{ff - - f}$$

$$({}^{i}{}_{j}) ( ) {}^{ij}$$

$$ff - = f - ff$$

$$_{i\,j\,j\,i\,ij\,i\,j}$$

$$f_{i}$$

d'où

$\{ ( )$ [redacted] $\} 0\ {}^{i}{}_{i\,j\,j\,ij\,j\,i} \sum \sum ff -$ [redacted] f [redacted] $\sum$ [redacted] $i \in I =$

étant donné que

$\{\}_{ijj}\sum f\,i\,I\in = f$, et $\{\}\,1_{\,i}\sum f\,i\in = I$

ii)- La distance entre points de $(N_J i)$

$'_{2'}1^{2}(,)\{\}^{i\,i}$

$d\,i\,i_{\,}j\,J$

$f = \in$

$j \qquad jj$

Cette distance mesure les proximités de forme entre lignes (ou entre colonnes) compte tenu de leurs poids différents. Elle est appelée dis[2] tance du χ (chi-2) et vérifie ce qu'on appelle

le

principe d'équivalence distributionnelle :

Si deux lignes (ou deux colonnes) du tableau $_{IJ}k$ *ont p oportionnelles et qu'on les remplace*

$s\ r$

par une seule ligne (ou par une seule colonne) qui n soit la somm colonne pa e e r

colonne

(ou la somme ligne par ligne), les distances entre colonnes (ou entre lignes) ne sont pas changées au sein du nuage (N(J) (ou N(I)).

En effet, si l'on considère deux éléments $_1j$ et $_2j$ de j tels que leurs profils sur j soient identiques ($^{1\,2}_{jj}$

$_{I\,I}f = f$) ; si on substitue aux colonnes $_1j$ et $_2j$ une colonne $_3j$ telle que : $_{s\,1\,2\,ij\,ij\,ij}f = + ff$, $_{s\,1\,2\,jjj}f = +$ ff, alors la distance entre éléments de I n'est pas modifiée. En d'autres termes on ne modifie pratiquement pas les résultats d'une analyse des correspondances si on regroupe deux rubriques très voisines en ajoutant leurs poids. $i\,(\,)\,N\,I_{\,J}$ ***iii)- La distance d'un point au centre de gravité du nuage est :*** $_{2\,2}{}^{1}(\,)\{\}$

$_i\rho\,i = \sum j \in J$

$_jf \qquad ^{jj}$

iv) De même on peut calculer l'inertie de ce point caractérisé par son profil i

$^{i}_{j}f$ ***et par son poids*** $_if$ ***.***

On a :

$^2(\,)(\,)_{\,ni}I\,i\,f = \rho\,i$

v) L'inertie du nuage $(\,)$ ***sera égale à*** $N\,I_{\,j}$

$(\,(\,)\{(\,)\}_{\,njn}\quad INI = \in \sum I\,i\,i\,I$

$$= {}^{1}{}_{2}\{\blacksquare\}^{ij} \qquad f$$

$$ff \sum_{f_{\blacksquare}^{i}} \sum_{ij} \blacksquare \in \qquad i\,I \qquad j\,i$$

$$fff = \sum\sum \blacksquare \in \in \qquad f^{fff}{}_{i\,I\,j\,J}$$

$$^{2}\{\blacksquare, \}^{ij\,i\,j}{}_{j\,i\,i}$$

d'où

$$^{1}{}_{I\,N\,I_{\blacksquare}} \qquad i\,I\,j\,J$$

$$ff = \blacksquare \sum\sum \in \in$$

$$^{2}(\,(\,))\,\{\blacksquare, \}_{nJij\,i\,j} \qquad _{ij}$$

compte tenu de la symétrie entre les indices i et j cette formule donc aussi l'inertie du nuage des points j .

On a donc :

$$(\,(\,))\,(\,(\,))_{nJn\,I}\,I\,N\,I = I\,N\,J$$

Remarque : Les profils peuvent être considérés comme de coordonnées euclidiennes. Si l'on considère la transformation suivante :

$\forall \in j\,J$, $\forall \in i\,I$, on associe à ${}^{i}_{j}f$ la quantité $_{jj}$

$$^{i\,1/2}ff \text{ alors la distance euclidienne usuelle}$$

entre deux points i et $'$

i vaut :

$$^{''}{}_{2\,'\,1/2\,1/2\,2}{}^{1}(\,,)\,(\,)^{2}$$

$$-- = - \sum_{)^{i\,i\,i}} \sum = - ,$$

$$d\,i\,i\,fffff_{f} \qquad (\,{}^{i}_{j}f$$

$$_{jjjjj} \qquad _{j}$$

et on voit qu'elle coïncide bien avec la distance du chi-2. Avec cette transformation le centre de gravité du n uveau nuage que l'on e $^{'\,1/2}(\,)\,\{(\,,)\,{}^{i}NI_{jjj\,i}fffi^{-} = \in I\}$ est :

o not

$$^{'\,1/2}\{\,;{}_{jj}f = \in fj\,J\}$$

De tout ce qui précède, on est conduit à diagonaliser la matrice des covariances, dont le T terme général s' crit :

é

$t = \sum$ $^{1/2\,1/2\,1/2\,1/2}( )(.)^{i\,i\,j\,j\,j}$

$_{jj}$ $^{j}{}_{jj}fffffff^{--}--;$,,

ce qui conduit à la ec

r herche des vecteurs propres et valeurs propres de la matrice des T'

v XX dans l'analyse générale.

ariances-covariances qui joue le rôle de

6.2.3. Analyse des points de $( ) NJ_{i}$ dans ^{n}R

Les ensembles I et J jouant un rôle parfaitement symétrique, l'analyse des points j de $( ) N J_{I}$ se déduit de l'analyse des points i de $N I( )$ par permutation des indices i et $_{J}j$ et des ensembles I et J .

6.2.4. Relations entre les points i de $( ) NI_{J}$ et les points j de $( ) NJ_{I}$ Co factoriels

sont deux à deux

mme en ACP, les facteurs sont de moyenne nulle et les axes

$( ) NI_{J}$ o

rthogonaux (au sens de la métrique du chi-2). On rappelle que l'inertie du nuage projeté sur l'axe α est égale à celle du nuage $( ) NJ_{I}$ projeté sur l'axe α (c'est la valeur propre de rang α). On a entre les éléments de I et de J les relations suivantes : $^{1/2}( ) (1/) \{ ( ) \}^{i} F i_{j} f$

$G j j J_{\alpha\alpha}$ $= \in \lambda \sum_{\alpha}$: projection de la ligne i sur l'axe de rang α de $( ) NI_{J}$; $^{1/2}( ) (1/) \{ ( ) \}^{i} G$

$j_{j} f F i i I_{\alpha\alpha}$ $= \in \lambda \sum_{\alpha}$: projection de la colonne j sur l'axe de rang α de $( ) N_{I} J$; λ_{α}: valeur commune de l'inertie associée à chacun de ces axes.

Ces formules sont appelées *formules de transition* et permettent la représentation simultanée des deux ensembles I et J et l'adjonction à l'un ou l'autre des deux ensembles supplémentaires de masse nulle. Cette expression d'une formule de transition est appelée propriété barycentrique : les éléments « lourds » attirant le barycentre, une colonne j attire d'autant plus une ligne i que la valeur $_{ij}f$ est élevée. Sur les plans factoriels, les points éloignés de l'origine, retiennent pa ulièrement l'attention, car ce sont les profils

rtic

les plus différents du profil moyen.

rs du tableau initial en fonction des marges et des

Enfin, on peut recalculer les valeu

facteurs. En effet, connaissant les lois marginales $_{I}f$ et $_{J}f$, la suite des facteurs F_{α} et G_{α} jusqu'à l'ordre p , et les valeurs propres $_{1},..., \lambda \lambda_{p}$, on trouve que :

$^{1/2}(1 \{(1/) ( ) ( ) [1,]\}_{ij\,i\,j}$ $fffF i G j p = +\sum \lambda \alpha_{\alpha\alpha\alpha} \in ;$

c'est la formule de *reconstitution* du tableau des données de départ.

6.2.5- Eléments supplémentaires

Soit $_{s}i$ une ligne supplémentaire. Pour visualiser $_{s}i$ sur le α − *ème* axe factoriel on projette le profil de $_{s}i$ sur cet axe. L'abscisse $( ) F i_{\alpha s}$ de cette proposition s'écrit : $_{1/2\,s}{}^{i}( ) (1/) \{ ( ) \} F i_{s j} f G j j J_{\alpha\alpha} = \in \lambda \sum_{\alpha}$

De même pour une colonne supplémentaire $_{s}j$, l'abscisse $( ) G j_{\alpha s}$ de la projection du profil $_{s}j$ sur l'axe α s'écrit : $^{1/2}( ) (1/) \{ ( ) \} {}^{sj} G j_{s i} f F i i I_{\alpha\alpha} = \in \lambda \sum_{\alpha}$

6.3- Interprétation d'une analyse factorielle des correspondances. 6.3.1. Tableau des données de base

Re tre b e prenons no exemple de le ta leau1. Le chef de la scolarité du CASP décide alors d mettre en place un système pour repérer les ts en fonc profil de le s étudian tion du urs note dans les différentes matières concernées. Le fich ser est donc un tableau ù ier à analy o chaque étudiant représente une ligne et chaq ère une co ue mati lonne.

6.3.2. Vecteurs et valeurs propres.

NUMERO ! VAL PROPRE 1 ! VAL PROPRE 2 ! VAL PROPRE 3 ! VAL PROPRE 4 ! ! 1.00000 ! .00750 ! .00515 ! .00327 ! OBJET 1 ! -.22264 ! .12390 ! -.02777 ! .07969 ! OBJET 2 ! -.27437 ! .15444 ! .14066 ! .00110 ! OBJET 3 ! -.24338 ! .12723 ! -.11714 ! .35144 ! OBJET 4 ! -.25169 ! -.05533 ! -.05474 ! -.06180 ! OBJET 5 ! -.23169 ! -.12409 ! -.77690 ! -.42761 ! OBJET 6 ! -.27993 ! .03318 ! -.03768 ! .02703 ! OBJET 7 ! -.21750 ! -.25530 ! .08272 ! .13412 ! OBJET 8 ! -.21272 ! -.70959 ! .33829 ! -.10346 ! OBJET 9 ! -.23826 ! -.18690 ! -.05908 ! -.03532 ! OBJET 10 ! -.23125 ! -.03079 ! -.09632 ! -.03212 ! OBJET 11 ! -.22402 ! .52647 ! .33496 ! -.46833 ! OBJET 12 ! -.21464 ! -.05100 ! .21780 ! -.49950 ! OBJET 13 ! -.21844 ! .12983 ! -.05285 ! .07794 ! OBJET 14 ! -.25735 ! .03713 ! -.01092 ! .26574 ! OBJET 15 ! -.25815 ! .13427 ! -.14161 ! .26705 ! OBJET 16 ! -.26638 ! .06293 ! .15184 ! .14257 ! OBJET 17 ! -.26444 ! -.02175 ! .13425 ! .12087 ! Tableau 8 : Vecteurs et valeurs propres de l'AFC.

En analyse factorielle des correspondances, toutes les valeurs propres sont comprises entre 0 et 1. En effet, on extrait *p* valeurs propres, avec $p \le \lceil \inf(cardI, cardJ) - 1 \rceil \lfloor$; on a : $_{1\,2}1 \ldots \ge \ge \lambda \lambda \ge \ge \lambda_{p}$.

Le vecteur propre associé à la valeur propre 1, est dénommé «vecteur propre trivial» car il n'apporte rien pour l'analyse factorielle de $( ) NIJ$ et $( ) NJI$. La première valeur propre à considérer dans notre exemple est en l'occurrence $_1 \lambda = 0.00750$. On a ensuite $_2 \lambda = 0.00515$ et $_3 \lambda = 0.00327$; chacune des 3 colonnes correspond à un vecteur propre (i.e les coordonnées des axes factoriels dans l'espace des 17 variables). Comme en ACP, l'histogramme (tableau 9) représente les valeurs
p
ropres par des lon urs qui leu ce qui perm
gue rs sont proportionnelles, et d'apprécier d'un regard la décroissance des valeurs propres ente. Sur notre exemple, on voit quand leur rang augm que chacune des deux premières valeurs propres est nettement λ_1 vaut près de 1.5
séparée de celle qui la suit :

fois λ_2 ; 1.6 $\lambda_{2\ 3} = \lambda$. En règle générale, une valeur propre bien séparée de celle qui la précède et de celle qui la suit est le signe que l'axe qui lui correspond est bien individualisé, et l'on cherchera à l'interpréter cet axe ; deux valeurs propres voisines l'une de l'autre, mais bien séparées des autres, sont le signe que le plan des axes qui leur correspond est bien individualisé. On rappelle enfi
n que
des valeurs propres élevées ‘’indiquent des oppositions tranchées dont l'interprétati
on est souvent à
la fois évidente et attendue. Des valeurs propres faibles peuvent correspond discrètes que l'analyse aura révélées’’.

LES VALEURS PROPRES VAL (1) = 1.00000
re à des corrélations plus

!NUM ! VAL PROPRE ! POURC.! CUMUL! VARIAT. ! *! HISTOG
RAMME DES VALEURS PROPRES

! 2 ! .00750 ! 27.273 ! 27.273 ! ******* ! * ! **** ********** ! ********** !

! 3 ! .00515 ! 18.744 ! 46.016 ! 8.529 ! * ! *** ********** ! ******

! 4 ! .00327 ! 11.889 ! 57.906 ! 6.854 ! * ! ** ! 5 ! .00289 ! 10.499 ! 68.405 ! 1.390 ! * ! ** ********** **********

! 6 ! .00254 ! 9.236 ! 77.641 ! 1.263 ! * ! ** *********

! 7 ! .00202 ! 7.340 ! 84.981 ! .1.836 ! * ! *** *****

! 8 ! .00105 ! 3.831 ! 88.812 ! 3.510 ! * ! * ***

! 9 ! .00086 ! 3.146 ! 91.958 ! .6 ! * ! *** 85

! 10 ! .00069 ! 2.5 03 ! 94.461 ! .643 ! * ! ***

! 52 ! 96.212 ! .751 ! * ! **
11 ! .00048 ! 1.7

! 12 ! .00039 ! 1.419 ! 97.632 ! .332 ! * ! **

! 13 ! .00027 ! .979 ! 98.610 ! .441 ! * ! *

! 14 ! .00017 ! .624 ! 99.234 ! .355 ! * ! *

! 15 ! .00010 ! .375 ! 99.609 ! .249 ! * !

! 16 ! .00007 ! .266 ! 99.875 ! .109 ! * !

! 17 ! .00003 ! .125 ! 100.000 ! .140 ! * !

Tableau 9 : Histogramme des valeurs propres

6.3.3 Les tableaux des facteurs sur I et sur J : aides à l'interprétation. Dans les tableaux ci-dessous (tableaux 10 et 11) et pour chaque colonne (comme en ACP) on rappelle les notions suivantes :

i) ***Le poids (POID)*** qui donne pour chaque étudiant i (ou pour chaque matière j) la part qu'il a dans le total du tableau. Le total de la colonne poids pour chacun des tableaux vaut 1000.

$/_{i\,i}f$ = $k\,k$; $_{j}/_{j}f$ = $k\,k$

ii) L'inertie (INR) qui donne en millièmes la valeur de l'inertie de chaque point $^{i}{}_{J}f$ de $(\,)\,N\,I_{J}$ (profil de la ligne afférente à l'étudiant *i)* ou $^{j}{}_{I}f$ de $(\,)\,N\,J_{I}$ (profil de la colonne afférente à la note j) par rapport au centre de gravité du nuage, rapporté à l'inertie totale du nuage.

$^{2\,2}(\,)\,(\,,\,)\,(\,)_{i}\,INR\,i$ = = $\underset{i}{f}\,d\,i\,G\,f\,\rho\,i$

! I ! QLT POID INR ! 1#F COR CTR! 2#F COR CTR ! 3#F COR CTR! 1 ! ABDO ! 398 41 15! -38 151 8! -24 59 5! -7 5 1! 2 ! BANZ ! 614 34 27! -20 19 2! 33 51 7! 8 3 1! 3 ! BATA ! 208 39 15! 32 100 5! -22 46 4! -3 1 0! 4 ! BOUK ! 562 44 40! 37 53 8! 41 66 14! -87 300 101! 5 ! BOYE ! 870 36 64! 108 241 56! -136 378 128! 107 236 126! 6 ! GOYI ! 869 42 31! -78 298 34! -39 75 12! 50 123 32! 7 ! LIK1 ! 983 38 53! 118 363 70! 123 391 110! 61 98 43! 8 ! LIK2 ! 787 42 29! -46 114 12! 3 0 0! 3 0 0! 9 ! LOUZ ! 839 36 77! -191 631 177! 90 138 56! -35 21 14! 10 ! MAKI ! 720 40 29! 51 134 14! -34 57 9! 13 9 2! 11 ! MALO ! 703 41 45! -91 272 45! -105 366 88! -26 23 9! 12 ! MAMP ! 693 42 40! -85 268 40! -20 15 3! 13 7 2! 13 ! MATO ! 772 41 49! -73 159 29! 103 321 84! 68 140 58! 14 ! MBIK ! 891 38 61! -184 773 172! 8 1 0! 20 9 5! 15 ! MPOU ! 789 40 26! 9 4 0! -19 19 3! 68 260 58! 16 ! NGUI ! 609 39 25! 42 98 9! -24 34 5! 59 197 41! 17 ! NKOK ! 699 41 17! 58 299 19! 39 131 12! 6 3 0! 18 ! NSEM ! 748 41 44! -7 1 0! 108 405 94! -62 134 49! 19 ! NSON ! 731 43 32! 1 0 0! -31 47 8! -84 341 93! 20 ! NZAK ! 918 42 76! 161 520 145! -26 13 5! -130 337 215! 21! ONDZ ! 984 38 54! 118 357 71! 122 378 109! 67 113 52! 22 ! SAFO ! 654 41 21! -7 3 0! 73 371 42! -43 128 23! 23 ! SAM1 ! 719 39 45! 12 5 1! -91 266 64! -73 170 64! 24 ! SAM2 ! 755 39 46 ! -48 73 12! -129 516 125! 17 9 3! 25 ! TSIB ! 668 42 40 ! 112 476 70! -37 52 11! 25 23 8! ! ! 1000 ! 1000 ! 1000! 1000! Tableau 10 : Facteurs sur I

iii) La qualité de représentation (QLT) qui s'interprète comme le carré du cosinus de l'angle que fait un point avec sa projection sur l'espace factoriel engendré par les axes factoriels : plus le cosinus est élevé, plus le point est corrél représenté sur cet axe.

On a ensuite, pour chaque facteur un groupe de trois colonnes.
é avec l'axe et donc bien

iv) Le facteur lui-mêm
e ; on sait que chaque point du plan est défini par ses deux

coordonnées ou facteurs. Seuls sont imprimés ici les trois premiers facteurs (1#F, 2#F et 3#F) sur les q extraits.

cin

COR

v) qui mesure la qualité de la représentation d'un point par sa projection sur l'axe. La somme des COR est égale à QLT.

$^{2}\cos(\ )/(\ )Fi_{i} = \rho i$

vi) La c CTR ***)***

ontribution relative d'un point à un axe (permet de repérer les éléments qui font l'axe, c'est-à-dire les éléments qui ont le plus de part à l'inertie du

n

uage projeté sur l'axe.

$^{2}(R_{i}CTfFi)/ = \lambda_{\alpha}$

! J1 ! QLT POID INR ! 1#F COR CTR! 2#F COR CTR! 3#F COR CTR! 1! STAT ! 465 50 32! 48 129 15! -9 4 1! 20 23 6! 2! MSTA ! 598 75 25! 49 260 24! 37 148 20! 0 0 0! 3! STAS ! 537 59 77! 45 57 16! -35 33 14! 83 190 124! 4! STAE ! 290 63 19! -19 44 3! -16 30 3! -14 24 4! 5! MATH ! 994 54 148! -46 28 15! -241 763 604! -106 147 183! 6! PROB ! 481 78 25! 10 12 1! -10 11 1! 6 4 1! 7! ECON ! 664 47 55! -102 325 65! 27 23 7! 35 39 18! 8! DEMO ! 957 45 173! -289 793 504! 114 124 114! -28 7 11! 9! INFO ! 623 57 30! -68 319 35! -18 22 3! -8 5 1! 10! GEOE ! 652 53 56! -12 5 1! -30 31 9! -8 2 1! 11! COME ! 964 50 143! 203 528 277! 107 147 112! -120 182 219! 12! COMN ! 880 46 70! -21 10 3! 73 126 47! -133 422 249! 13! TEXP ! 621 48 50! 51 93 17! -17 11 3! 20 15 6! 14! ANGL ! 560 66 25! 12 15 1! -3 1 0! 59 331 71! 15! HIST ! 576 67 41! 45 120 18! -39 92 20! 59 207 71! 16! STAG ! 601 71 15! 20 73 4! 41 292 23! 31 163 20! 17! APPG ! 368 70 16! -7 8 0! 36 216 18! 26 111 15! ! ! 1000! 1000! 1000! 1000!

Tableau 11 : Facteurs sur J

En examinant ces tableaux on peut repérer :

les points excentrés pour lesquels INR est nettement supérieur à $POID$ comme LOUZ, -

NZAK (pour les individus) et DEMO, COME origine ; l'

et peut être dû qu'à leur distance élevée à MATH (pour les variables) ; cela ne

les points centraux, pour lesquels $POID$ est nettement supérieur à INR - comme ABDO, BANZ (pour les individus), ST et APPG (pour les variables) ;

AG

les points correspondant aux CTR les plus forts. L'interprétation des axes reposera sur -

l'examen de ces points qui ''font'' l'axe. Dans notre exemple on sélectionnera pour le premier axe variables) ; DEMO et COME (pour les

LOUZ, MBIK et NZAK (pour les individus),

- des points au centre de gravité : ce sont ceux pour les axes qui expliquent l'écart

lesquels COR a une forte valeur ;

- LIK1 et BOYE pour les individus et Les points bien représentés comme ONDZ,

MATH, COME et DEMO pour les variables.

6.3.4. Représentations graphiques

En AFC, on utilise la représentation simultanée de $(\ )\ NI_J$ et $(\ )\ NJ_I$ sur les plans de coordonnées, rapportés chacun à deux axes factoriels. On sait, d'après les formules de transition (cf §6.4.2) que, au coefficient $\lambda^{-1/2}$ près, les points représentatifs d'un nuage sont sur un axe, les barycentres des points représentatifs de l'autre. On constate que ce coefficient est supérieur à 1, et le nuage est d'autant plus dilaté dans la direction d'un axe que la valeur propre correspondante est faible.

Deux points de $(\ )\ NI_J$ proches révèlent un comportement semblable des caractères lignes correspondant pour c les points de $(\ )\ NJ_I$. pour les proximités entre

es deux axes de projection (il est de même

L'interprétation des proximités entre les projections des points de $(\ )\ NI_J$ et de $(\ )\ NJ_I$ est plus délicate ; ''le seul cas dans lequel on puiss tenir comp la proximité entre les

e te de

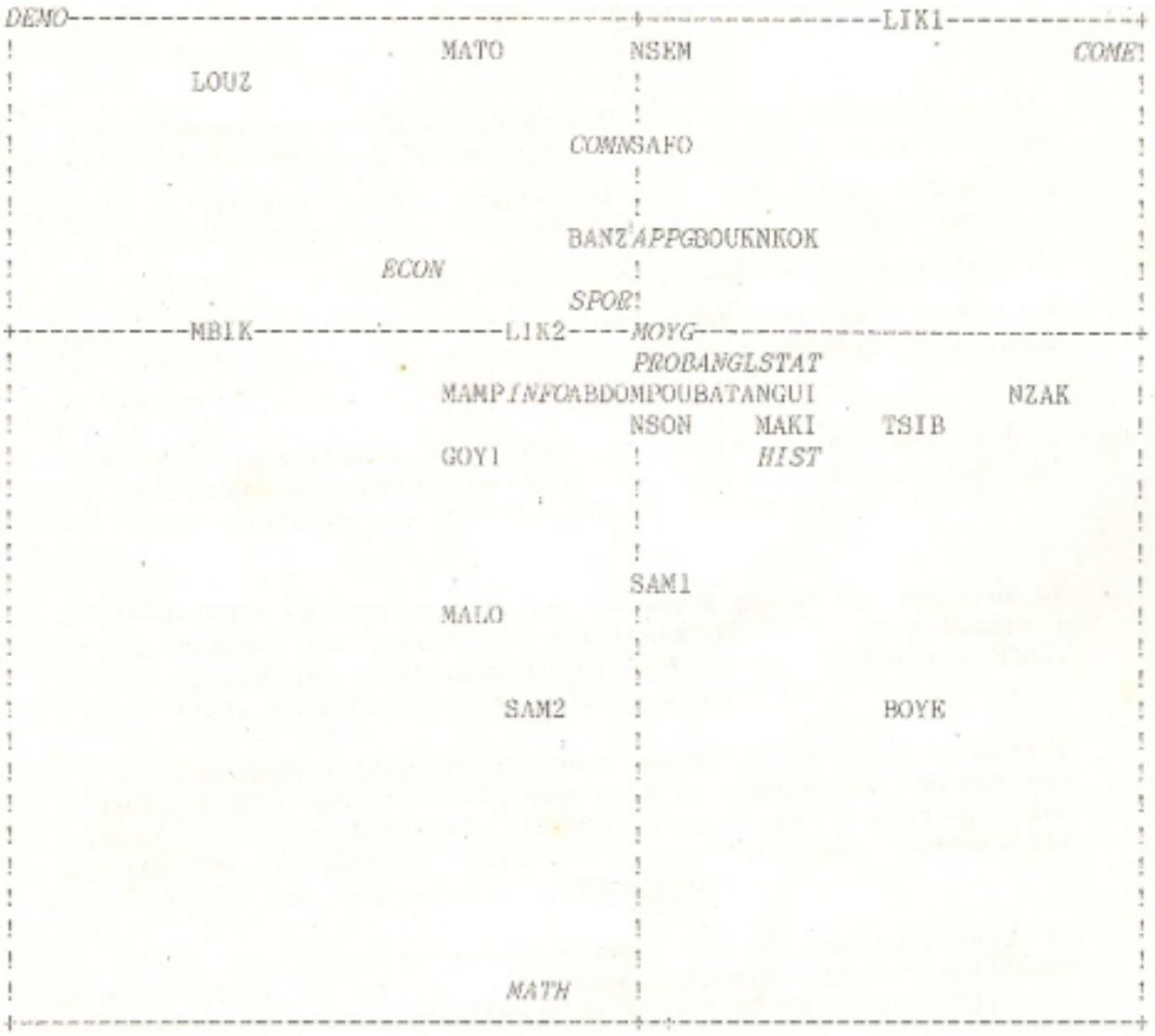
projections de deux points appartenant l'un à $(\ )\ NI_J$, l'autre à $(\ )\ NJ_I$, est celui où ces deux points sont situés à la périphérie du nuage… Lorsqu'il s'agit par contre des po s situ

int és à

l'intérieur du nuage, les proximités sont un véritable piège pour l'intuition''[60]. Dans notre exemple, par rapport aux facteurs imprimés (ici 3) les représentations possibles sont les plans (1,2), (1,3) et (2,3). Nous n'examinerons que les plans (1,2) et (1,3) (graphiques 6 et 7).

AXE HORIZONTAL (1) – AXE VERTICAL (2). Nombre de points : 44

ECHELLE : 4 caractères = 0.27 1 ligne = .011



Nombre de points superposés : 7 ONDZ(LIK1) MSTA(NKOK) STAG(BOUK) STAE(ABDO) TEXP(NGUI) STAS(MAKI) GEOE(NSON)

Graphique 6 : Représentation simultanée des individus et des variables dans l'espace factoriel (1,2).

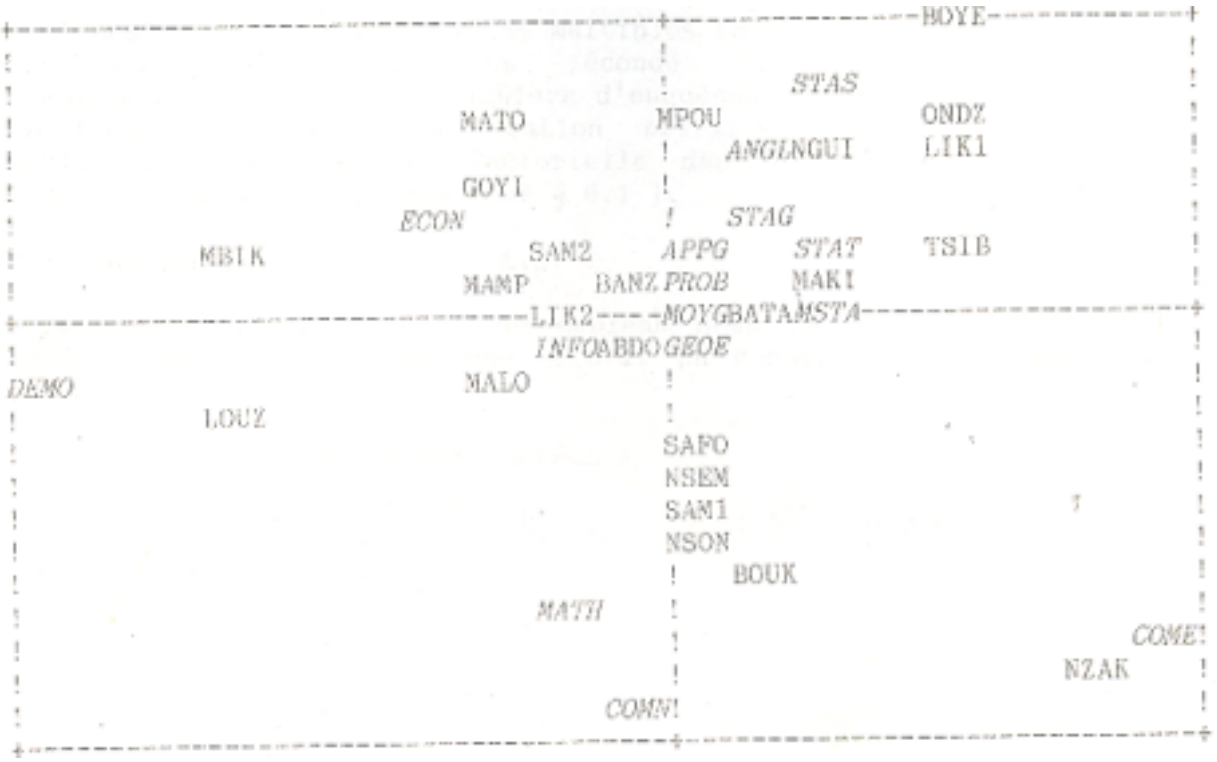
En prenant en compte tous les éléments ci-dessus énumérés on peut rapidement interpréter les résultats de la façon suivante : l'axe 1 oppose les individus LOUZ et MBIK aux individus NZAK, TSIB, et LIK1, la variable DEMO à la variable COME. En combinant l'analyse des deux ensembles, on peut constater que LOUZ et MBIK qui ont une bonne note en démographie, et ont une mauvaise note en comptabilité d'entreprise. Par contre, les individus NZAK, TSIB et LIK1 qui sont bons en comptabilité d'entreprise, sont médiocres en démographie.

L'axe 2 peut être considéré comme l'axe des mathématiques. En bas et à gauche de cet

axe, on trouve effectivement les meilleurs étudiants dans cette discipline : ce sont SAM2, MALO et SAM1.

Sur l'axe 3 les oppositions ne sont pas très tranchées. Cet axe est néanmoins dominé par COME. On trouve ce point à la périphérie du nuage à côté de NZAK ; cette proximité peut être expliquée par le fait que la meilleure note en comptabilité d'entreprise a été obtenue par cet étudiant.

AXE HORIZONTAL (1) – AXE VERTICAL (3). Nombre de points : 44
ECHELLE : 4 caractères = 0.27 1 ligne = .011



Nombre de points superposes:5HIST(NGUI) TEXP(STAT) NKOK(MAKI) SPOR(BANZ) STAE(ABDO)

Graphique 7 : Représentation des individus et des variables dans l'espace factoriel (1,3).

Note : Le lecteur pet être tenté de comparer les résultats graphiques issus d'une ACP de ceux d'une AFC. En effet, la tentation est grande, car on peut considérer l'analyse factorielle des correspondances comme une ACP classique sur des données transformées (profils) utilisant une distance particulière, la distance du chi-2. Si l'on tente cette comparaison écrit PH Cibois ''on s'aperçoit que les résultats sont comparables à cette différence qu'en analyse en composantes principales, seules les lignes et colonnes les plus fortes en effectif sont prises en compte ce qui n'est pas le cas en analyse des correspondances où une pondération est introduite''[17]. En tenant compte du fait que, le premier facteur extrait en analyse factorielle des correspondances est un facteur trivial (cf.

§6.3.2), λ =1 et que l'on ne numérote pas, il poursuit : ''si l'on veut donc comparer les facteurs équivalents et en prenant la numérotation usuelle des facteurs il faut donc comparer le facteur 2 de l'analyse en composantes principales avec le facteur 1 de l'analyse des correspondances et ainsi de suite… Il est d'ailleurs un cas où les deux méthodes donnent les mêmes résultats graphiques, c'est dans la situation où les marges du tableau sont identiques et où de ce fait, la pondération de l'AFC ne modifie rien''.

6.4 L'analyse des correspondances multiples

L'analyse des correspondances multiples (ACM), est considérée comme l'application la plus féconde de l'analyse des correspondances et dont les fichiers d'enquêtes socio-économiques constituent le champ d'application privilégié ; elle est l'application de l'analyse factorielle des correspondances à l'étude des tableaux logiques (cf §6.1).

6.4.1- Tableau disjonctif complet

Reprenons la description du tableau sous forme disjonctive complète commencée au paragraphe §6.1. en conservant les mêmes notations.

Les marges du tableau sont notées :

$\{\}_{i\,ij\,ij\,q}\ k = \in \sum$ [redacted] $k\,j\,J =$ [redacted] $k\,j \in j\ q \in Q = CardQ$ = nombre de questions avec $\{\}_{ij\,q}\ 1 \sum k\,j \in = j$

$\{_{j\,ij}\ k\,k = \sum\ j \in J\}$ = nombre d'individus ayant fourni la réponse j à la question q .

donc

$\{\}_{j\,q} \sum k\,j \in = J\ CardI = n$

{[redacted]}$_{ij\,i}\ k\,k = \sum \sum$ [redacted] $i \in I = nCardQ$

$_1J = 2\ _2J = 4\ _3J = 5\ _4J = 3$

i

1 0	0 0 1 0	0 0 0 0 1	0 1 0	$_{i}k\ C=$
			$_{j}k$	$k\ n=$

On note les fréquences marginales comme suit :

$/ / 1/ _{i\,i}f = = k\ k\ CardQ\ nCardQ = CardI$

Soit $_{j}p$ la proportion des individus ayant fourni la réponse j à la question q : $/ P$

$k_{j\,j} = CardI\ .\ (\{\}\ 1)_{j\,q} \sum p\ j \in = J\ .$

On a alors : $/ _{j\,j}f = p\ CardQ$

Le nombre d'individus ayant fourni la réponse j est $_{j}I \cdot p$; cela veut dire que le chiffre 1 se trouve $_{j}I \cdot p$ fois dans la colonne j ; et le chiffre 0 se trouve $I\ (1\ )_{j} \cdot - p$ fois. Le tableau précédent peut être interprété de la manière suivante : dans une enquête comprenant quatre questions, l'individu a choisi la modalité 1 de la première question i (question à deux modalités de réponses), la modalité 3 de la deuxième question (question à quatre modalités de réponses), la modalité 5 de la troisième question (question à cinq modalités de réponses) et la modalité 2 de la quatrième question (question à trois k modalités de réponses). On vérifie bien que la ligne de marge $_{i}$ est égale au nombre de questions c'est-à-dire 4. Le tableau ainsi construit est formé par la juxtaposition de 4 tableaux logiques et contient autant de fois la valeur 1 qu'il y a de ces tableaux (ici 4 bien sûr).

Les tableaux disjonctifs complets ont le défaut d'être grands et leur analyse coûteuse : une variante consiste à effectuer l'analyse factorielle des correspondances sur ''le tableau de Burt''.

6.4.2- Tableau de Burt

Si l'on croise l'ensemble des modalités du tableau disjonctif complet avec elles-mêmes, on obtient le tableau de Burt.

$$B = \in \sum k\ k\ i\ I$$

$\forall \in j\ J\ ,\ ' \forall \in j\ J : {}_{\cdot\cdot}\{\}^{ij}{}_{jj\,ij}$

= nombre d'individus ayant adopté à la fois les modalités j et $'j$.

$$'\{\}^{qj}{}_{jj}\sum B\,j\,J\in = k$$

$$B = \in \sum B\,j\,J = k\;CardQ$$

$$'\{\}^{jj}{}_{jj}$$

$$''\{,\}\,(_{jj}{}^{2}B = \in \sum B\,j\,J\,j \in J = n\;CardQ)$$

$_1J$ $_2J$ $_4J$		$_3J$		
0 0				
	0 0			
		0 0		$B_{jj} = k\;CardQ$
			0 0	
				$^2n\;C(\;ardQ)$

$_1J$

$_2J$

$_3J$

$_4J$

Il faut faire remarquer que : si j et $'j$ appartiennent au même sous-ensemble de

$_{q}j$

modalités, on a :

$', _{q}j\, j$: $\forall \in J\ 'j \neq j$

$B_{,}$ = 0si $_{jj}$

= $_{j}k$ si $'j = j$ (nombre d'individus ayant adopté la modalité j). Le tableau de Burt ainsi construit, est donc formé d'une juxtaposition des tableaux de contingence entre les variables prises deux à deux. Les tableaux contenant la diagonale croisent chaque variable avec elle même et sont remplis de 0 à l'exception de leurs diagonales, remplies des effectifs de chaque modalité.

6.4.3- Equivalence entre les deux analyses précédentes

On considère les deux tableaux k et $_{IJ}B_{JJ}$définis par :

$\{ , _{IJ\,ij}k\, k = \in i\, I\, j \in J\}$ et

$'\{ , \}_{jj}$

B_{JJ}= $\in B_{'}j\, J\, j \in J$

avec

$_{\cdot ij}{}_{jj\,ij}\{ /.\}\ B_{i} = \in \sum k\, k\, k\, i\, I$

On a alors :

$B = = \sum\sum B \sum k\, k\, k = \sum k\, k \sum k = \sum k = k$

$/\ (\ /\ )^{j\, ij\, i\, ij\, i}{}_{jj\,ij\,ij}$

$'''$

$'''$

$B = = \sum\sum B\, B = \sum k$ $'$

$_{jj\,i\,i\,j\,i}$ $_{ij\,j}$

$= k$

= − B B B B B B) ....

En tenant compte du fait que

$$B = k ; B k = :$$

$$B_{jj} = k ; {}''_{jj}$$

$$u\,k \quad \lfloor \rfloor \sum \lceil \rceil \; k\,k\,k\,k\,k\,k\,k$$

$$^{1/2}\{ / \} / /( )^{ij\,i\,j\,j}{}_{jj\,j\,j}$$

$$,,,{}_{ij} = -$$

$u\,t$ = = rme général de la matrice de var nc T

On a donc ${}''_{jj\,jj}$ te ia es-covariances du §6.2.2.

Ainsi l'analyse factorielle de B_{JJ} revient à diagonaliser la matrice ${}^{2}T$. L'analyse des correspondances de ${}_{IJ}k$ fournit les mêmes facteurs que celle du tableau B_{JJ} mais, les valeurs propres correspondantes sont différentes : à la valeur propre λ de analyse de correspond la valeur propre ${}_{IJ}k^{2}\lambda$ de l'analyse de B_{JJ} l' .

6.4.4- Calculs de contributions dans l'analyse du tableau disjonctif complet.

i) Le carré de la distance au centre de gravité d'un point j ***s'écrit dans*** ${}^{n}R$ ***:*** ${}^{2\,2}( ) / {}^{2}($

$$,) \{1 ((/)) \}_{\,i\,ij\,j\,i} \quad \rho\,j = d\,j\,G = \sum fff - f\,i \in I$$

Comme : $1/{}_{i}f$ = $CardI$ et $/{}_{jj}f = p\;CardI$

on a en définitive :

$$^{2\,2}\;^{1}( )\{(\blacksquare \}_{ij}$$

$$CardQ_{\,j\;CardI\,\blacksquare\,i\;I}$$

$$p\;CardI\;\rho = \blacksquare \sum^{\in}$$

$$_{j}$$

On peut décomposer cette somme selon les valeurs prises par ${}_{ij}f$; on trouve que

$$: \quad ^{1}( )_{ij\,j}{}^{f\;CardI\;p}\;CardQ\;CardI$$

$$= \cdot$$

. fois et 0 $(1)\;Ca_{\,j}rdI \cdot - p$ fois ; ce qui donne :

$$^{2}( )^{2\,1}( )(1)^{j}\,p$$

$$\rho\,j\,p\,p^{\;-} = +{}_{j} - {}_{j}$$

$$_{j}p$$

$$^{2}( )(1)_{jj}\,\rho\,j = -\,p\,p$$

ii) La contribution de la modalité j ***vaut donc :***

$$^{2\,2}( )(,)(/)(,)\;CTR_{\,jj}j = = f\,d\,j\;G\;p\;CardQ\;d\,j\;G$$

$$( )(1)/\;CTR_{\,j}j = - p\;CardQ$$

L'inertie due à une modalité est d'autant plus gra nde que l'effectif dans cette modalité est faible. On évitera de définir les modalités que l'on peut supposer a priori trop rares.

iii)- La c ontribution d'une question q est :

$$CTR_{jq}(q) = \{(1)/\} - \sum_{i \in j} p\ CardQ$$

Comme $\sum_{j \in q} p_j = \{\} 1_{jq}$ on a :

$$CTR_q(q) = -Cardj\ CardQ\ (1/)\ (1)$$

Elle est proportionnelle au nombre de modalités de la question. Du point de vue du codage des données cela suppose que, le nombre de modalités de chaque question doit être voisin pour avoir des poids équivalents pour chaque question.

iv)- L'inertie totale est égale à :

$$NJ = -CardJ\ CardQ\ (1/)\ (1)\ (I_n)$$

On remarque que cette inertie ne dépend pas des liaisons existant entre les variables. Elle vaut 1 si boutes les questions ont deux modalités de réponse.

analyse des correspond ances multiples.

6.4.5- Interprétation d'une truire un tableau

6.4.5.1- Tableau des données de base.

A partir du tableau des variables quantitatives (tableau1), on peut cons de description logique. La procédure est la suivante :

- on rend toutes les variables qualitatives par découpage en classes. Le découpage peut se faire soit en classes d'effectifs égaux, soit en classes d'amplitudes égales ; - la connaissance du domaine à étudier peut aussi conduire l'utilisateur à fixer lui-même les bornes de classes ;

- dans tous les cas, il est conseillé avant tout découpage, de construire les histogrammes des variables pour l'ensemble des individus. Ces derniers sont une aide précieuse pour délimiter les bo nes des classes.

r

De tout ce qui précède, on retiendra tout simplement ‘'qu'en analyse de vos propres données aussi il faudrait «parfois » considérer la statistique comme une science expérimentale"[ ] 31 . Dans l'exemple choisi, on a découpé chaque variable-matière en trois classes d'amplitudes égales. On a 19 · 3 = 57 variables nouvelles issues des 19

variables d'origine. Ainsi par exemple, pour la variable STAT, on aura les trois modalités STA1, STA2, et STA3. Le tableau disjonctif complet $_{IJ}k$ associé au découpage précédent est donc un tableau 25 · 57. Cependant, on sait que l'ACM est très sensible aux modalités rares qui peuvent perturber l'analyse (i.e rendre instable les axes) et reléguer sur des axes ultérieurs des phénomènes plus intéressants. On peut donc provisoirement abandonner ces modalités et par la suite les positionner en éléments supplémentaires[49]. Fort de ce qui

p ariabl né taires. Ce sont récède, six v es seront position es en éléments supplémen : PROB, ANGL, HIST, SPOR, A tab

PPG et MOYG. Le leau des variables actives est donc de dimension 25 · 39 et celui des variables supplémentaires de dimension 25 · 18.

6.4.5.2. Valeurs propres

En A M

C chaque valeur propre est inférieure ou égale à 1 et leur somme est égale à l'inertie totale du nuage, soit : = dans notre exemple. Dans $(CardJ - = CardQ)/CardQ\ (39 - 13)/13\ 2$ ces conditions, aucune valeur propre ne peut représenter plus que 100/inertie totale, soit : 100

$$CardQ\ /(CardJ - CardQ)$$

LES VALEURS PROPRES VAL (1) = 1.00000

NUM ! VAL PROPRE ! POURC ! CUMUL ! VARIAT. ! ! HISTOGRAMME DES VALEURS PROPRES ! 2 ! .26448 ! 13.224 ! 13.224 ! ******* ! * ! ********** ! ********** ! 3 ! .23987 ! 11.994 ! 25.218 ! 1.230 ! * ! ********** ! ********** 4 ! .21365 ! 10.683 ! 35.900 ! 1.311 ! * ! ********** ! ********* 5 ! .17759 ! 8.880 ! 44.780 ! 1.803 ! * ! ********** ! ***** 6 ! .16227 ! 8.113 ! 52.893 ! .766 ! * ! ********** ! *** 7 ! .14655 ! 7.327 ! 60.220 ! .786 ! * ! ********** ! ** 8 ! .13548 ! 6.774 ! 66.994 ! .553 ! * ! ********** ! 9 ! .12249 ! 6.124 ! 73.119 ! .650 ! * ! ********** 10 ! .08959 ! 4.479 ! 77.598 ! 1.645 ! * ! **********

11 ! .07595 ! 3.797 ! 81.395 ! .682 ! * ! *********

12 ! .06882 ! 3.441 ! 84.836 ! .356 ! * ! ********

13 ! .06724 ! 3.362 ! 88.198 ! .079 ! * ! ********

14 ! .05017 ! 2.509 ! 90.707 ! .854 ! * ! ******

15 ! .04701 ! 2.351 ! 93.057 ! .158 ! * ! *****

16 ! .03601 ! 1.801 ! 94.858 ! .550 ! * ! ****

17 ! .03142 ! 1.571 ! 96.429 ! .229 ! * ! ****

18 ! .01971 ! .985 ! 97.414 ! .586 ! * ! **

19 ! .01881 ! .940 ! 98.355 ! .045 ! * ! **

20 ! .01462 ! .731 ! 99.086 ! .209 ! * ! **

21 ! .00950 ! .475 ! 99.560 ! .256 ! * ! *

22 ! .00530 ! .265 ! 99.825 ! .210 ! * ! *

23 ! .00309 ! .155 ! 99.980 ! .110 ! * !

24 ! .00040 ! .020 ! 100.000 ! .135 ! * !

25 ! .00000 ! .000 ! 100.000 ! .020 ! * !

26 ! .00000 ! .000 ! 100.000 ! .000 ! * !

27 ! .00000 ! .000 ! 100.000 ! .000 ! * !

28 ! .00000 ! .000 ! 100.000 ! .000 ! * !

29 ! .00000 ! .000 ! 100.000 ! .000 ! * !

30 ! .00000 ! .000 ! 100.000 ! .000 ! * !

31 ! .00000 ! .000 ! 100.000 ! .000 ! * !

32 ! .00000 ! .000 ! 100.000 ! .000 ! * !

33 ! .00000 ! .000 ! 100.000 ! .000 ! * !

34 ! .00000 ! .000 ! 100.000 ! .000 ! * !

35 ! .00000 ! .000 ! 100.000 ! .000 ! * !

36 ! .00000 ! .000 ! 100.000 ! .000 ! * !

37 ! .00000 ! .000 ! 100.000 ! .000 ! * !

38 ! .00000 ! .000 ! 100.000 ! .000 ! * !

39 ! .00000 ! .000 ! 100.000 ! .000 ! * !

2

Tableau 12 : Histogramme des valeurs propres de l'ACM

On peut dire que les valeurs propres issues d'une ACM sont donc un peu particulières et difficilement interprétables (taux d'inertie faibles). Elles donnent une idée très pessimiste de l'information extraite[43].

6.4.5.3 Tableaux des facteurs sur I et sur J

Les coordonnées factorielles des points i de $NI(\ )$ et j de sont données par les $NJ(\ )$ mêmes formules que celles de l'AFC ainsi que les résultats numériques et les paramètres associés. Cependant on peut faire constater les résultats suivants dans le tableau des J facteurs sur :

- toutes les questions ont le même poids soit 1/13=0.077 (77 en millième) ; la somme des poids des modalités d'une même variable vaut donc (en millième) 77 ; - les questions ayant le même nombre de modalités $3\ Ca_{q}rdJ$ = , les contributions à l'inertie de toutes les questions sont égales : $T(\ )(\ 1)/\ 2\ /13\ 0.154\ CR_{q}q$ = $CardJ -$ = $CardQ$ =

! I1 ! QLT POID INR ! 1#F COR CTR! 2#F COR CTR! 3#F COR CTR! 1! ABDO ! 734 40 56 ! -602 130 55 ! 649 151 70 ! -1127 454 238 ! 2! BANZ ! 393 40 31 ! 512 169 40 ! -318 65 17 ! 496 159 46 ! 3! BATA ! 42 40 31 ! -172 19 4 ! -164 18 5 ! -88 5 1 ! 4! BOUK ! 425 40 53 ! 543 111 45 ! 757 215 96 ! 511 98 49 ! 5! BOYE ! 344 40 29 ! 324 72 16 ! -587 236 57 ! 231 36 10 ! 6! GOYI ! 39 40 40 ! 13 0 0 ! 15 0 0 ! 282 39 15 ! 7! LIK1 ! 813 40 38 ! -1097 628 182 ! -412 88 28 ! 429 96 34 ! 8! LIK2 ! 310 40 41 ! 771 287 90 ! 195 18 6 ! -99 5 2 ! 9! LOUZ ! 233 40 37 ! 355 69 19 ! -543 160 49 ! 83 4 1 ! 10! MAKI ! 118

40 37 ! -214 24 7 ! 414 92 29 ! -64 2 1 ! 11! MALO ! 433 40 39 ! 699 253 74 ! -239 30 10 ! -540 151 55 ! 12! MAMP ! 406 40 43 ! 59 2 1 ! -615 177 63 ! -696 227 91 ! 13! MATO ! 376 40 46 ! 80 3 1 ! -713 222 85 ! -588 151 65 ! 14 ! MBIK ! 445 40 38 ! 705 263 75 ! -515 140 44 ! -280 42 15 ! 15 ! MPOU ! 113 40 37 ! -47 1 0 ! 316 54 17 ! 326 58 20 ! 16 ! NGUI ! 77 40 34 ! -289 50 13 ! -175 18 5 ! 119 8 3 ! 17! NKOK ! 224 40 35 ! 186 20 5 ! 355 72 21 ! 482 132 43 ! 18! NSEM ! 165 40 39 ! 505 130 39 ! -177 16 5 ! 195 19 7 ! 19! NSON ! 348 40 42 ! -852 346 110 ! 63 2 1 ! -1 0 0 ! 20! NZAK ! 503 40 42 ! 48 1 0 ! 710 242 84 ! 735 260 101 ! 21! ONDZ ! 813 40 38 ! -1097 628 182 ! -412 88 28 ! 429 96 34 ! 22! SAFO ! 135 40 41 ! 57 2 0 ! 336 56 19 ! 396 77 29 ! 23! SAM1 ! 473 40 57 ! -271 26 11 ! 849 253 120 ! -744 194 104 ! 24! SAM2 ! 269 40 35 ! -415 100 26 ! -539 168 48 ! -50 1 0 ! 25 ! TSIB ! 375 40 42 ! 199 19 6 ! 747 265 93 ! -437 91 36 ! 1000 ! 1000 ! 1000 ! 1000 !

Tableau 13 : Facteurs sur I

L'inertie d'une question $INR(\ ) q$ la question q ramenée à l'inertie totale vaut donc : $(\ )(\ )_q$ $INR\ q = CTR\ j$ /inertie totale = 0,154/2=0,077 (77 en millième). La somme des inerties des modalités d'une même variable vaut (en millième) 77.

! J1 ! QLT POID INR ! 1#F COR CTR ! 2#F COR CTR ! 3#F COR CTR ! 1 ! STA1 ! 388 34 22 ! 553 240 39 ! -213 36 6 ! 378 112 23 ! 2 ! STA2 ! 493 25 26 ! -856 345 68 ! -71 2 1 ! -556 145 36 ! 3 ! STA3 ! 80 18 29 ! 129 5 1 ! 485 74 18 ! 48 1 0 ! 4 ! MST1 ! 777 40 18 ! 470 240 33 ! -687 511 79 ! -156 26 5 ! 5 ! MST2 ! 561 15 31 ! -69 1 0 ! 1355 459 118 ! -636 101 29 ! 6 ! MST3 ! 516 22 28 ! -824 264 55 ! 308 37 9 ! 744 215 56 ! 7 ! STS1 ! 426 34 22 ! 516 209 34 ! 220 38 7 ! 478 180 36 ! 8 ! STS2 ! 555 25 26 ! -1002 472 93 ! -309 45 10 ! 285 38 9 ! 9 ! STS3 ! 547 18 29 ! 391 48 11 ! 9 0 0 ! -1257 499 137 ! 10! STE1 ! 453 34 22 ! -62 3 0 ! -749 441 79 ! -111 10 2 ! 11! STE2 ! 386 37 20 ! 199 36 6 ! 432 172 29 ! 439 178 33 ! 12! STE3 ! 622 6 35 ! -849 63 17 ! 1530 204 60 ! -2023 356 118 ! 13! MAT1 ! 243 28 25 ! -118 8 1 ! -355 71 15 ! 540 164 38 ! 14! MAT2 ! 107 28 25 ! 406 93 17 ! 156 14 3 ! 42 1 0 ! 15! MAT3 ! 297 22 28 ! -371 53 11 ! 256 26 6 ! -749 218 57 ! 16! ECO1 ! 158 28 25 ! 368 76 14 ! 379 81 17 ! 54 2 0 ! 17! ECO2 ! 416 31 23 ! -700 327 57 ! -351 82 16 ! -100 7 1 ! 18! ECO3 ! 122 18 29 ! 615 120 26 ! 18 0 0 ! 85 2 1 ! 19! DEM1 ! 190 34 22 ! -331 86 14 ! 187 27 5 ! 312 77 15 ! 20! DEM2 ! 65 31 23 ! 131 11 2 ! 281 53 10 ! -23 0 0 ! 21! DEM3 ! 469 12 32 ! 582 65 16 ! -1218 282 76 ! -801 122 37 ! 22! INF1 ! 264 43 17 ! -303 116 15 ! -204 53 7 ! 273 95 15 ! 23! INF2 ! 282 22 28 ! 288 32 7 ! 442 76 18 ! -669 174 45 ! 24! INF3 ! 68 12 32 ! 555 59 14 ! -60 1 0 ! 216 9 3 ! 25! GEO1 ! 335 28 25 ! -410 95 18 ! -646 235 48 ! 97 5 1 ! 26! GEO2 ! 475 25 26 ! 780 286 57 ! 599 169 37 ! 208 20 5 ! 27! GEO3 ! 103 25 26 ! -318 48 9 ! 128 8 2 ! -317 47 12 ! 28! COE1 ! 372 28 25 ! 630 223 42 ! -502 142 29 ! 108 7 2 ! 29! COE2 ! 265 34 22 ! -82 5 1 ! 280 62 11 ! -502 198 40 ! 30! COE3 ! 455 15 31 ! -954 228 53 ! 288 21 5 ! 910 207 60 ! 31! CON1 ! 500 28 25 ! -735 304 57 ! -532 159 33 ! 257 37 9 ! 32! CON2 ! 384 31 23 ! 187 23 4 ! 172 20 4 ! -715 341 74 ! 33! CON3 ! 485 18 29 ! 790 197 44 ! 512 83 20 ! 806 205 56 ! 34! TEX1 ! 455 28 25 ! -884 440 82 ! -39 1 0 ! -162 15 3 ! 35! TEX2 ! 627 28 25 ! 727 297 55 ! -713 286 59 ! -279 44 10 ! 36! TEX3 ! 504 22 28 ! 202 16 3 ! 967 363 84 ! 567 125 32 ! 37! STG1 ! 202 31 23 ! 334 74 13 ! -433 125 24 ! -67 3 1 ! 38! STG2 ! 277 22 28 ! -303 36 7 ! 788 241 56 ! -4 0 0 ! 39! STG3 ! 25 25 26 ! -152 11 2 ! -148 10 2 ! 87 4 1 ! 1000 ! 1000 ! 1000 ! 1000 !

Tableau 14 : Facteurs sur J

6.4.5.4- Représentation graphique

Essayons d'interpréter le plan (1,3). Essayer parce que l'intérêt d'une telle étude est relativement limitée. Parce que aussi ces méthodes ont été conçues pour l'analyse de très grands tableaux. En soumettant ce tableau (vu son format) à une analyse de

correspondances multiples, nous avons voulu avant tout, favoriser le côté pédagogique. La spécificité ici, contrairement aux méthodes précédentes, réside dans le fait que l'étude ne porte plus sur les variables elles-mêmes, mais sur les modalités de ces variables. On réalise une analyse par niveaux de variable, plus poussée que celles des variables initiales.

L'axe 1 (graphique 8) : les variables les plus contributives à la formation de cet axe sont GEO2, TEX2, STS2, TEX1, STA2, ECON2 et CON1. Cet axe est donc essentiellement dominé par les modalités 2 et 1 de ces variables. Du côté positif de l'axe on trouve des variables GEO2 et TEX2 ; du côté négatif, les variables STS2, TEX1, STA2, ECO2 et CON1. On peut dire que l'axe 1 oppose entre elles les variables dont les modalités ont des valeurs moyennes (modalités 2).

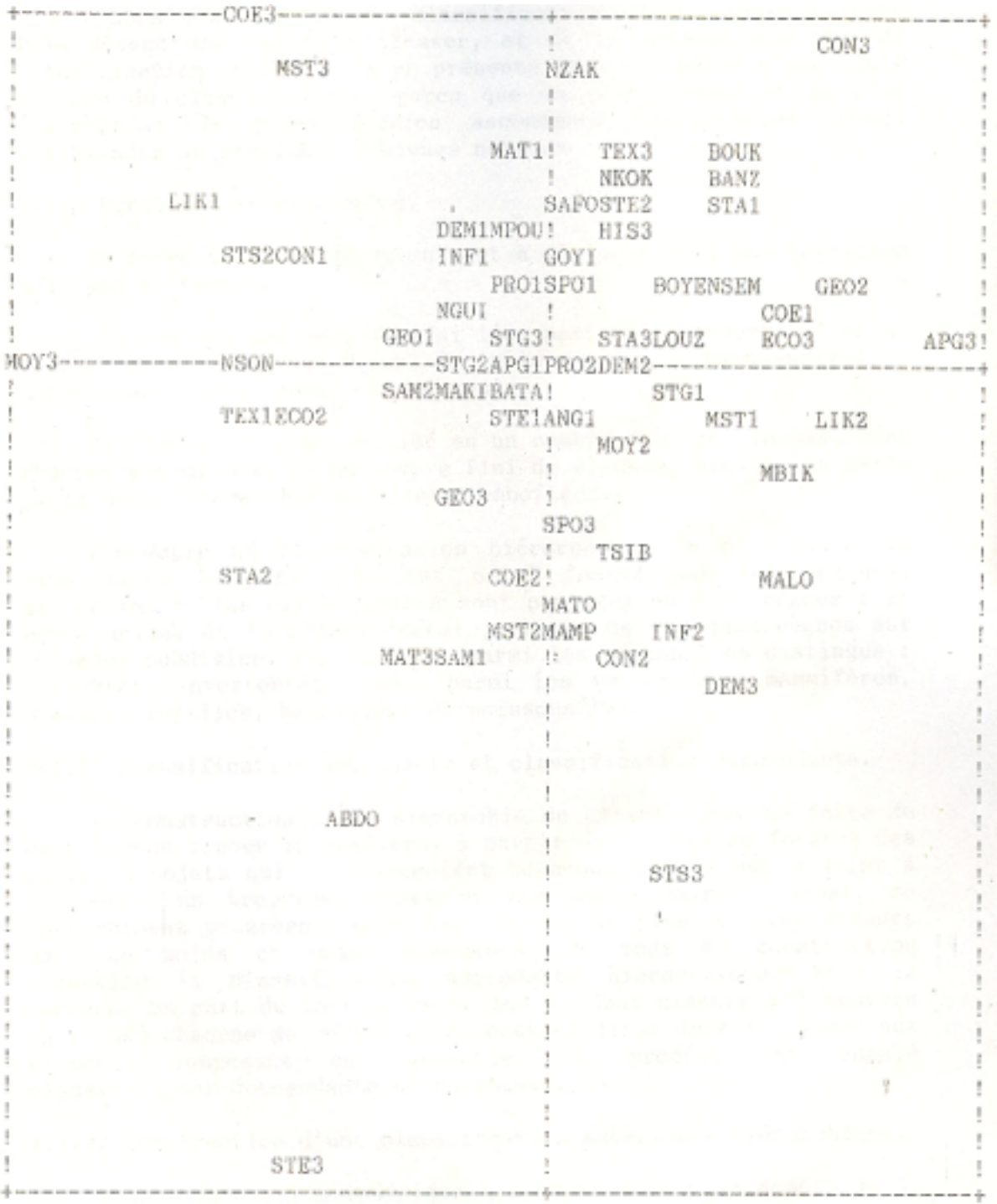
En ce qui concerne les individus, l'axe 1 oppose les individus MBIK, LIK2, BOUK et NSEM aux individus LIK1, ONDZ, NSON et ABDO. Si on raisonne en termes de groupe d'individus, on trouve du côté positif de l'axe, les étudiants ayant un niveau moyen en culture générale (GEO2, TEX2) et du côté négatif, les étudiants moyens dans certaines disciplines comme la statistique et l'économie et ayant obtenu des mauvais résultats dans d'autres matières telles que la comptabilité ou les techniques d'expression.

L'axe 3 est dominé par les valeurs élevées des modalités de certaines variables. Ce sont, du côté positif de l'axe MST3, CON3 et COE3 et du côté négatif STS3, STE3 et MAT3. L'opposition de ces variables sur cet axe constitue le trait dominant. On peut, comme pour l'axe 1, tirer les conclusions similaires en ce qui concerne les individus ou groupe d'individus.

Bien que ne participant pas à la formation des axes, on peut utilement interpréter les différentes positions des variables supplémentaires sur les axes (comme les positions respectives de MOY3 et APG3 sur l'axe 1).

Enfin, signalons que pour rendre l'interprétation plus facile, il est conseillé de joindre par un trait les modalités successives d'une même variable. Cela est très important surtout lorsque les modalités des variables sont assez nombreuses (cas des enquêtes).

AXE HORIZONTAL (1) – AXE VERTICAL (3). Nombre de points : 82
Echelle : 4 caractères = .161 1 ligne = .067



Graphique 8 : Représentation des individus et des variables dans l'espace factoriel (1,3)

7. La classification ascendante hiérarchique

7.1. Principes généraux

Disons d'emblée qu'une classification n'est jamais unique. Elle dépend des objets à classer, et de la méthode pratique de classification utilisée. On ne présente dans ce cahier qu'une méthode de classification, parce que la plus connue et la plus éprouvée : *la classification ascendante hiérarchique (CAH).* Définissons au préalable quelques notions.

7.1.1. Partition et hiérarchie

Le terme de classification sert à désigner soit une partition soit une hiérarchie. On obtient une *partition* si l'on partage un ensemble I en un système de classes non vides, de telle sorte que tout individu i appartienne à une classe et une seule. Si l'ensemble I est divisé en un nombre fini de classes, dont chacune est divisée en un nombre fini de classes, etc., on

parle alors d'une *hiérarchie de classes* sification hiérarchique le
emboîtées. L'exemple de clas

plus connu et sans doute le plus cité est celui fourni par les sciences naturelles : ''les êtres vivants sont partagés en deux règnes : le règne animal et le règne végétal ; chacun de ces deux règnes est lui-même subdivisé. Par exemple, parmi les animaux, on distingue : vertébrés, invertébrés ; puis parmi les vertébrés : mammifères, oiseaux, reptiles batraciens

,

et [5].
poisson
s''

7.1.2. Classification ascendante et classification descendante

La construction d'une hiérarchie de classes peut se faire de deux façons : pour la première, à partir de la base en formant des paires d'objets qui se ressemblent beaucoup ; puis on adjoint à une paire un troisième objet ou une autre paire ; ainsi, se construisent progressivement des classes de plus en plus grandes mais de moins en moins homogènes. Ce mode de construction s'appelle la classification ascendante hiérarchique. Pour la seconde, on part du tout qu'on scinde en deux classes ; à nouveau on scinde chacune de celles-ci en deux et ainsi de suite jusqu'aux éléments composant cet ensemble. Ce procédé est appelé classification descendante hiérarchique.

7.1.3- Construction d'une classification ascendante hiérarchique. Supposons que les individus à classer soient au nombre de 5. C'est-à-dire que I ={1,2,3,4,5}. On construit la classification ascendante hiérarchique (CAH), à partir d'un ensemble J , selon le processus suivant :

- on calcule les distances entre les individus pris deux à deux.
- on choisit un critère qui permet d'agréger les différents éléments pour former des classes.

Après avoir défini la distance et le critère d'agrégation, le processus se poursuit selon les étapes suivantes (voir schéma ci-dessous) :

- on cherche les deux éléments de I les plus proches. Sur la figure, ce sont par exemple les éléments (4) et (2) que l'on agrège en un seul élément noté (6). Cet nouvel élément est appelé *nœud.* Il est défini par ses deux successeurs : *l'aîné* et *le benjamin* (éléments (4) et (2)), son poids (nombre d'éléments) et son *indice de niveau* (ici le nombre ν_6) qui n'est autre que la distance entre les éléments ((4) et (2)).
- selon le même critère d'agrégation choisi, on calcule les distances entre le nouvel élément (ici (6)) et les éléments restants. On se retrouve dans les conditions de l'étape précédente, mais cette fois-ci, avec 4 éléments seulement à classer.
- on renouvelle le processus jusqu'à ce qu'il n'y ait plus qu'un seul élément (élément (9)).

niveau

V_9 V_8

V_7

(6)

V_6 0

(9)

(7)

(8)

(4) (2) (1) (5) (3)

Par rapport à l'exemple ci-dessus mentionné on peut faire le récapitulatif suivant :

I = {1, 2,3, 4,5} est l'ensemble des éléments sur lesquels est édifié la classification ;

N = {6,7,8,9} est l'ensemble des nœuds, ou des classes construites ; ce sont :

; ;

6 { = 4, 2} 7 { = 5,3} 9 { = 1, 2,3, 4,5}

Si l'on note respectivement par $A(\ )\ n$ et $B(\ )\ n$ l'aîné et le benjamin on a :

; ; ;

$A(9) = 8$ $A(8) = 6$ $A(7) = 5$ $A(6) = 4$

$B(9) = 7$ $B(8) = 1$ $B(7) = 3$ $B(6) = 2$

; ; ;

L'ensemble des classes terminales de la classification est l'ensemble de ses éléments minimaux (composés de classes réduites à un élément) : $T = \{(1),(2),(3),(4),(5)\}$. Les éléments terminaux sont numérotés de 1 à $CardI - 1$. Les nœuds de la classification sont numérotés de $CardI + 1$ à $2\ 1\ CardI -$.

7.1.3. Critères d'agrégation

La construction de la CAH dépend de la formule choisie pour le critère d'agrégation, ce qui revient à définir une distance entre classes. On expose ici quatre critères classiques, en insistant sur l'un d'entre eux : le critère de l'inertie que l'on adoptera dans la suite. ***i)- Le critère du saut minimum (d_{saut}) :***

$q\ ' q\ (\ ' q\ q, q$

entre les ensembles de points et d_{saut}) est la distance minima entre un point de q' et un point de . Le critère du saut minimum consiste donc à choisir la plus petite des distances qui permet de passer d'une classe à une autre.

ii)- Le critère du diamètre (d_{diam}) :

($'q\ q, q\ 'q$

d_{diam}) est la distance maxima entre un point de et un point de . On prend pour distance entre les classes, la plus grande de toutes les distances.

iii)- Le critère de la distance moyenne (d_{moy}) :

($'q\ q, q\ 'q$

d_{moy}) est la moyenne des distances entre un point de et un point de . Ce critère apparaît comme un compromis des deux critères précédents.

iv)- Le critère selon la variance (ou critère de l'inertie) :

pour le calcul de ce critère, on suppose que I est considéré comme un nuage de l'ensemble

points munis de masse dans un espace euclidien. C'est justement le cas de l'exemple traité dans ce cahier où les étudiants sont repérés en fonction du profil de leurs notes. Ce tableau peut donc être considéré comme un tableau de contingence ou comme un tableau de mesures).

Soit I un ensemble fini ; ( ) le nuage des éléments de $N\ I_J\ i\ I$ et affectés de masse . On m_i

rappelle qu'une inertie est le produit d'une masse par le carré d'une distance.

$$(\ )\ N\ I_J$$

_ L'inertie du nuage s'écrit :

$$(\ (\ ))\ \{\ (\ )_{n\ J\ n}\ I\ N\ I = \in \sum I\ I\ i\ I\}$$

$$^2\{\ (\ )\ m\ I_i = \in \sum \rho\ i\ I\}$$

$^{2\ 2}\rho\ (\ )i\ d$ = $(i,G)\ G\ i$ où mesure le carré de la distance au centre de gravité du point .

Soit une partie de

$m_i\ G\ q$

$q\ I$, on notera par sa masse totale, et ou simplement son centre de gravité.

- L'inertie d'une classe s'écrit :

$$^2(\ )\ \{\ (\ ,\ )\ \}_{n\ i}\ I\ q\ m = \in \sum^{q} d\ i\ q\ i\ q\ ;$$

et l'inertie d' une partition Q de I sera égale à :

$$(\ )\ \{\ (\ )\ \}_{n\ n}\ I\ Q\ I = \in \sum q\ q\ Q$$

$$= \in \sum q\ G\ q\ Q\}\ ;\ \{\ \}\ m\ m_{q\ i} = \in \sum i\ q$$

$$^2\{\ (\ ,\ )\ m\ d_q$$

A toute partition de I en un ensemble de classes correspondant une décomposition $Q\ q$

$(\ )\ N\ I_J$

de l'inertie du nuage en inertie interclasses et inerties intra-classes suivant la formule

(relation de Huygens) :

$$I_{nJn}(Q) = I_N(Q) + \sum_{q \in Q} I_q(q)$$

L'inertie intra-classe est d'autant plus faible que les classes obtenues sont plus compactes ; et l'inertie interclasse est d'autant plus élevée que les classes de la partition sont bien séparées. En d'autres termes, l'inertie intra-classe est une bonne mesure de l'homogénéité d'une classe, de même l'inertie interclasse est une bonne mesure de la différence entre les classes. Soient maintenant deux classes a et b de Q_{n-1} respectivement de masse et

m_a m_b

n Q_n m $m_{a\,b}$ + . Le critère qui conduit aux que l'on agrège en une seule classe de de masse

a b

choix de et est celui qui rend minimum la perte d'inertie réalisée en passant de Q_{n-1} Q_n

à :

$$\nu_1(n) = [I(Q_{n-1}) - I(Q_n)]_{nnn} \text{ minimum.}$$

Q_n

Ce qui équivaut à maximiser l'inertie de la partition (ou encore à maximiser le moment centré d'ordre 2 de la partition). A chaque l'inertie intra-classe de la pas, on minimise

partition construite. La quantité m m $_{n\,d}$ +

$_{a\,b}$

n

$_2(n) = \nu(a, b)$

$\nu = a\ b$

$m\ m$

$\nu(n)$ est également appelée indice de niveau (cf.§ 7.1.3).

L'inertie totale de la classe peut selon la relation de Huygens être décomposée en inertie a

b n $\nu(n)$, terme proportionnel au carré de la

des deux classes et dont la réunion est et

distance entre aîné et benjamin du nœud n :

$$I(a\,b)_n = I_a + I_b + \nu(n)$$

Les $\nu(n)$ fournissent la décomposition totale du nuage. On a :

$$\sum_{n \in J} \nu(n) = I_N = I_J$$

7.2. Interprétation d'une classification ascendante hiérarchique.

7.2.1. Le tableau des données

Comme au §6.3.1, on considère le tableau 1 comme un tableau de correspondance et on désire édifier une CAH sur l'ensemble §6.2.1 :

I des étudiants. On rappelle quelques

formules du

- la masse de l'élément i de I : k ;

$$/ m f_{iii} = = k$$

- la distance de chi-2 entre profils :

$$d i i = - \sum fffj \in J$$
$$jjj \quad {}^{,2'2}(,) \{1/ ( ) \}^{ii}$$

- l'inertie de l'élément i :

$$I i f = - \sum fffj \in J$$
$$nijjj \quad {}^{2}( ) \{1/ ( ) \}^{i}$$

- la masse de la classe q :

$$\{ \}_{qi} f = \in \sum f i q$$

- le profil de la classe q :

$$\blacksquare f_{ff\blacksquare} = \blacksquare \sum$$
$$j_q{}^{q}{}_{ij} \quad j \quad qq$$

- l'inertie du centre de gravité de la classe q par rapport au centre de gravité du nuage : $_j$

$$\rho q f = - \sum ffj {}^{2}( ) ( )_{nq}$$
$$jj \quad {}^{22}( ) \{1/ ( ) \}^{q}$$

$I q f = \rho q$; avec $\in J$

- l'indice de niveau du nœud n :

$$ffn\blacksquare j_{ff}$$
$$+ \sum \in$$
$$v = \blacksquare J \quad {}^{22}( ) \blacksquare \}^{abab}$$
$$ab \quad jjj$$

7.2.2- Histogramme des indices de niveau de la hiérarchie

Chaque ligne donne successivement : le numéro j du nœud, l'indice de niveau $I( ) J$, exprimé en millième, les numéros de l'Aîné $A( ) J$ et du Benjamin $B( ) J$, le taux d'inertie $T J( )$ afférent au nœud qui est le rapport de l'inertie du nœud $I( ) J$, à l'inertie totale du nuage (exprimé en millième) et IN E DE NIVEAU le taux d'inertie cumulé $T Q( )$.

SOMME DIC S .29124 E-01 DES

J ! B(J)! T(J) ! T(Q) ! HISTOGRAMME DES INDICES DE NIVEAU ! I(J) ! A(J)!

! 49 ! 4 ! 46 ! 48 ! 150 ! 150 ! ********** ! 48 ! 3 ! 32 ! 47 ! 118 ! 268

! ********** ! 47 ! 3 ! 42 ! 45 ! 86 ! 354 ! **********

! 46 ! 2 ! 26 ! 43 ! 79 ! 433 ! **********

! 45 ! 2 ! 40 ! 44 ! 72 ! 505 ! **********

! 44 ! 2 ! 31 ! 39 ! 56 ! 560 ! **********

! 43 ! 2 ! 20 ! 41 ! 52 ! 612 ! **********

! 42 ! 1 ! 37 ! 34 ! 43 ! 655 ! **********

! 41 ! 1 ! 38 ! 33 ! 42 ! 697 ! **********

! 40 ! 1 ! 36 ! 27 ! 39 ! 736 ! **********

! 39 ! 1 ! 13 ! 35 ! 31 ! 767 ! *********

! 38 ! 1 ! 29 ! 19 ! 28 ! 795 ! ********

! 37 ! 1 ! 5 ! 16 ! 24 ! 819 ! *******

! 36 ! 1 ! 28 ! 25 ! 24 ! 843 ! *******

! 35 ! 1 ! 2 ! 30 ! 23 ! 866 ! *******

! 34 ! 1 ! 24 ! 23 ! 23 ! 889 ! *******

! 33 ! 1 ! 18 ! 4 ! 20 ! 909 ! *****

! 32 ! 1 ! 9 ! 14 ! 19 ! 928 ! *****

! 31 ! 1 ! 11 ! 12 ! 19 ! 947 ! *****

! 30 ! 0 ! 8 ! 1 ! 16 ! 963 ! ****

! 29 ! 0 ! 22 ! 17 ! 13 ! 976 ! ****

! 28 ! 0 ! 10 ! 3 ! 12 ! 988 ! ***

! 27 ! 0 ! 15 ! 6 ! 12 ! 1000 ! ***

! 26 ! 0 ! 21 ! 7 ! 0 ! 1000 ! *

Tableau 15 : Histogramme des indices de niveau

Av c re l t adop e, la s mme des indices de n ale à l'inertie ec le ritè de 'iner ie té o iveau est ég to ale d nua e des individus.

t u g

B(J)! P(J) ! DESCRIPTION DES CLASSES DE LA HIERARCHIE

! J ! I(J) ! A(J)!

! 49 ! 4 ! 46 ! 48 ! 25 !

! 48 ! 3 ! 32 ! 47 ! 17 ! LOUZ MBIK BOYE NGUI SAM2 SAM1 MAKI BATA TSIB MPOU GOYI MALO MAMP MATO BANZ LIK2 ABDO ! 47 ! 3 ! 42 ! 45 ! 15 ! BOYE NGUI SAM2 SAM1 MAKI BATA TSIB MPOU GOYI MALO MAMP MATO BANZ LIK2 ABDO

! 46 ! 2 ! 26 ! 43 ! 8 ! ONDZ LIK1 NZAK SAFO NKOK NSON NSEM BOUK ! 45 ! 2 ! 40 ! 44 ! 11 ! MAKI BATA TSIB MPOU GOYI MALO MAMP MATO BANZ LIK2 ABDO

! 44 ! 2 ! 31 ! 39 ! 6 ! MALO MAMP MATO BANZ LIK2 ABDO ! 43 ! 2 ! 20 ! 41 ! 6 ! NZAK SAFO

NKOK NSON NSEM BOUK ! 42 ! 1 ! 37 ! 34 ! 4 ! BOYE NGUI SAM2 SAM1

! 41 ! 1 ! 38 ! 33 ! 5 ! SAFO NKOK NSON NSEM BOUK

! 40 ! 1 ! 36 ! 27 ! 5 ! MAKI BATA TSIB MPOU GOYI

! 39 ! 1 ! 13 ! 35 ! 4 ! MATO BANZ LIK2 ABDO

! 38 ! 1 ! 29 ! 19 ! 3 ! SAFO NKOK NSON

! 37 ! 1 ! 5 ! 16 ! 2 ! BOYE NGUI

! 36 ! 1 ! 28 ! 25 ! 3 ! MAKI BATA TSIB

! 35 ! 1 ! 2 ! 30 ! 3 ! BANZ LIK2 ABDO

! 34 ! 1 ! 24 ! 23 ! 2 ! SAM2 SAM1

! 33 ! 1 ! 18 ! 4 ! 2 ! NSEM BOUK

! 32 ! 1 ! 9 ! 14 ! 2 ! LOUZ MBIK

! 31 ! 1 ! 11 ! 12 ! 2 ! MALO MAMP

! 30 ! 0 ! 8 ! 1 ! 2 ! LIK2 ABDO

! 29 ! 0 ! 22 ! 17 ! 2 ! SAFO NKOK

! 28 ! 0 ! 10 ! 3 ! 2 ! MAKI BATA

! 27 ! 0 ! 15 ! 6 2 ! MPOU GOYI

! 26 ! 0 ! 21 ! 7 2 ! ONDZ LIK1

Tableau 16 : Description des classes de la hiérarchie

L'histogramme des indices de niveau est édité pour permettre à l'utilisateur de voir comment varient les indices de niveau, et d'indiquer à quel niveau on peut couper l'arbre de classification pour avoir une partition convenable (classes stables). ''Si la décroissance e st très forte, ceci symbolise le fait qu'il n'existe que quelques séparations principales. Les niveaux les plus bas de la hiérarchie peuven de t être considérés comme des intermédiaires calcul comme cela se présente pour les axes de l'analyse des correspondances. On prendra cependant soin d'examiner des séparations à des niveaux faibles''.

7.2.3- Le tableau du contenu des classes

On a construit classes. Les classes de la hiérarchie J $CardI - 1$ classes, c'est-à-dire 25-1=24 sont numérotées de 26 à 49. Chaque classe est décrite par : son numéro , son indice de niveau nombre de ses éléments et la liste des $PJ( )$ $I( ) J$, ses successeurs $A( ) J$ et $B( ) J$, le

é
léments de chaque classe. Ces éléments sont rangés dans l'ordre où ils sont imprimés en marge de l'arbre. Prenons par exemple la classe 40 ; on a d'abord les trois individus de la classe 36 c'est-à-dire (MAKI, BATA et TSIB), puis les deux éléments de $A(40) = 36$ (MPOU et $B(40) = 27$ GOYI).

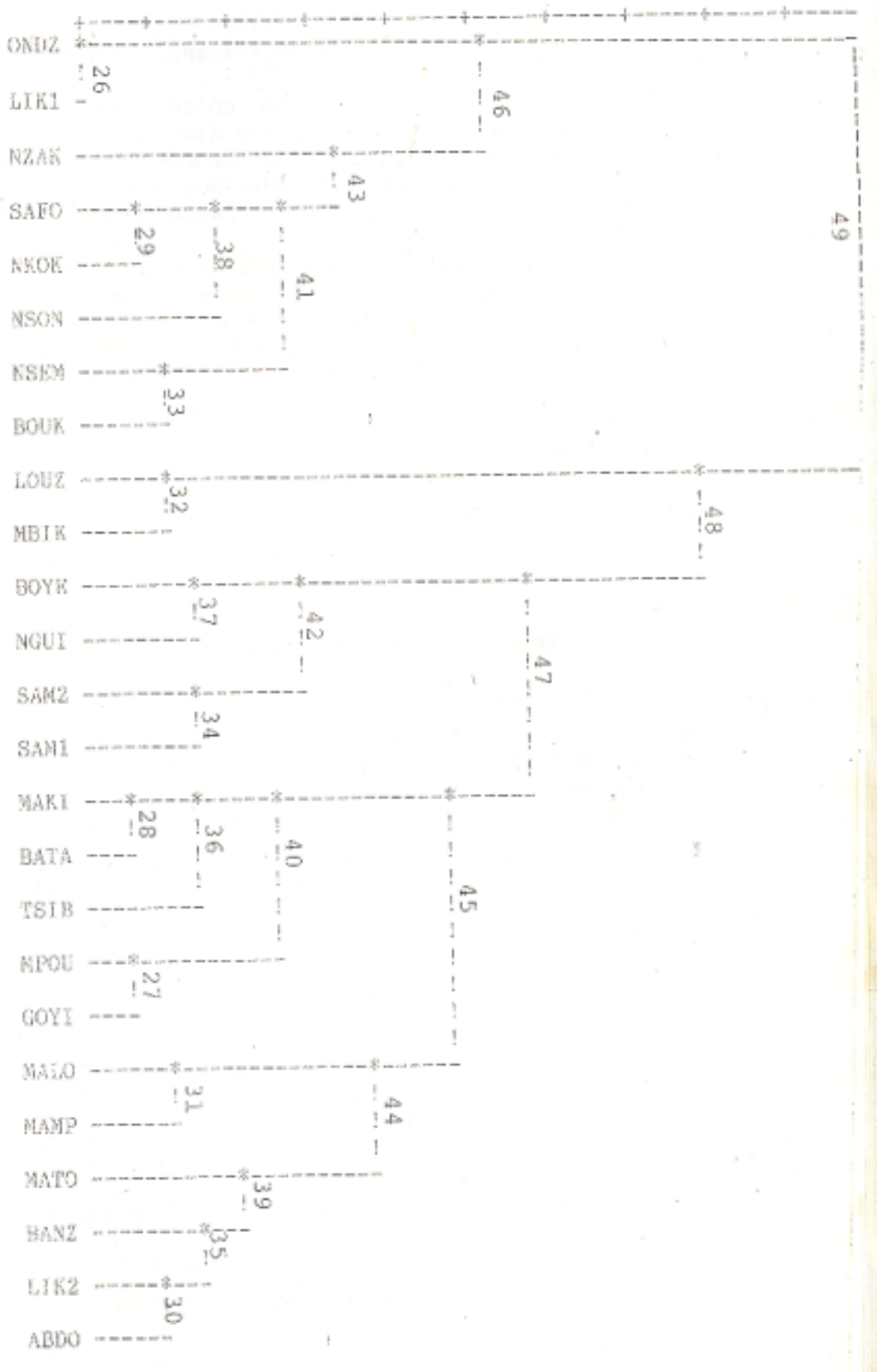
7.2.4. L'arbre de classification hiérarchique

Du tableau du contenu des classes, on déduit l'arbre de classification (Graphique 9) qui, comme on l'a déjà dit, définit un système emboîté de classes. La lecture descendante de l'arbre, dans le sens inverse de sa construction, permet d'examiner les partitions comprenant peu de classes. Si on coupe l'arbre au niveau le plus élevé, on obtient deux classes. En effet, en partant du sommet, le nœud 49 se scinde en ses deux successeurs immédiats $A(49) = 46$ et $B(49) = 48$.

Si on coupe maintenant l'arbre légèrement au dessus du niveau du nœud 48, on obtient une partition en trois classes. En coupant ensuite l'arbre entre les nœuds 46 et 47, on obtient une partition en quatre classes. De toute évidence, ces classes, seront d'autant plus nombreuses que la coupure de l'arbre sera proche des éléments terminaux.

''L'examen de l'arbre amène en fait le praticien à privilégier certaines partitions, jugées « bonnes », et à en rejeter d'autres, jugées « mauvaises »''[60].

On retrouve ici, à quelques exceptions près, la typologie suggérée par l'analyse factorielle des correspondances (cf. § 6.3.4). Les deux approches sont donc complémentaires et nous allons maintenant examiner leur interprétation conjointe.



Graphique 9 : Représentation de classification hiérarchique

7.2.5- Calcul de contributions

L'arbre de classification établi au §7.2.4 permet de distinguer les classes les unes des autres, mais ne permet pas de connaître comment ces classes se sont formées et comment elles se séparent. Les calculs des contributions sont effectués pour le savoir. L'objectif de ces calculs est de préciser :

- en quoi une classe s'écarte du centre de gravité du nuage : c'est l'étude de q l'excentricité des classes par rapport à des axes. Cette étude des classes se fait soit par rapport à des variables (axes initiaux des variables associés à un tableau de données), soit par rapport à des axes factoriels ;
- en quoi diffèrent les deux successeurs $a\ n(\ )$ et $b(n)\ n$ en lesquels se scinde la classe .

C omme précédemment, cette étude sera réalisée d'abord dans l'espace rapporté au système d'axes factoriels, ensuite dans l'espace rapporté au systèm variables ;

- les contributions mutuelles entre classes et facteurs.

des axes. Formulaire $_{IJ}k$

e des axes définis par les

7.2.5.1 Etude des classes par rapport à

Soit , un tableau de correspondance. On a effectué sur ce tableau une analyse factorielle des correspondances et une classification ascendante hiérarchique sur I. On a donc sur I, un ensemble de facteurs et un système de classes. Soit q une partie de I et on note par $_{q}f$ sa masse. Sur les axes de l'analyse factorielle des correspondances, on peut placer la classe q ; q étant un barycentre de points i, on a : $( )\ \{ ( )/ \}\ F\ q\ _{i\,q}fF\ i\,f\ i\ q\ _{\alpha\alpha}$ = $\sum$ ∈ , avec $\{ \}\ _{q\,i}f$ = ∈ $\sum$ $f\ i\ q$

Pour une classe , on calcule :

q

q α :

- le cosinus carré de l'angle formé entre le rayon vecteur de la classe et l'axe . C'est la contribution relative de l'axe $^{2\,2\,2}COR\ (\ )\ q\ Cos\ (\ )\ q\ F\ (\ )\ q\ /\ (q)\ _{\alpha\alpha}$ = = $_{\alpha}$ρ α à l'excentricité q^2 (($rho\ q$) = ρ $(q)\ q$

de la classe ,) étant l'excentricité de la classe par rapport au centre du nuage).

$COR\ (\ )\ q\ _{\alpha}$ ⇒ α intervient peu dans l'écart de au centre du nuage.

#0 q

$COR\ (\ )\ q\ _{\alpha}$ ⇒ α explique l'écart de au centre du nuage.

#1

q

$COR\ q$

La somme des donne la qualité de la représentation de la classe dans l'espace des facteurs retenus :

$\sum COR$ = QLT

q α : ^{2}( ) ( )/ CTR_q - la $_\alpha q$ = $f\ F_{\alpha\ \alpha}$ α λ .

contribution relative de la classe au facteur

CTR est la contribution relative de la classe q à l'inertie interclasse par rapport à l'inertie totale de l'axe α .

CTR ( ) q $_\alpha$#0 $\Rightarrow$ q n'explique pas l'inertie de l'axe α ;

CTR ( ) q $_\alpha$#1 $\Rightarrow$ α n'explique pas l'inertie de l'axe α .

7.2.5.2. Etude des classes par rapport à des axes. Exemple

a) Etude des classes par rapport à des axes factoriels

Les deux tableaux relatifs aux classes, concernent pour l'un, les cinq classes les plus hautes de la hiérarchie (tableau 17) et pour l'autre, les classes terminales de cette hiérarchie restreinte (tableau 18). Cette étude des classes est faite par rapport à des axes factoriels.

O

n a extrait cinq facteurs ; seuls les deux premiers facteurs sont indiqués. Pour le premier ^{2} tableau la première lign ρ ( ) q e n'est pas utilisable (sauf la colonne poids) puisque le est égal à zéro. Le poids total du nuage étant normalis poids 1000/1000=1.

(TOUTES VA R MULTIPLIEE R LES LEU S SONT S PA 1000)

é à 1, le centre du nuage 49 à pour

AXES FACTORIEL A
S 1 2

CLASSE AINE BNJMN ! POIDS INR QLT ! C CT F CO CT F 1 OR R ! 2 R R

49 46 48 ! 1000 0 0 ! 0 0 0 ! 0 0 0 48 32 47 ! 673 49 984 ! -29 401 71 ! -29 396 104 47 42 45 ! 599 47 945 ! -8 31 5 ! -38 630 160 46 26 43 ! 327 101 984 ! 60 401 146 ! 60 396 214 45 40 44 ! 446 42 548 ! -21 166 25 ! -17 101 23 ! ! 248 ! 501

Tableau 17 : Facteurs pour les 5 classes les plus hautes de la hiérarchie

Dans cette étude de la position des classes par rapport aux axes factoriels, on note que le premier facteur est moyennement corrélé avec la classe 48 et 46 (0.401) ; le deuxième

facteur est corrélé avec la classe 47 (0.630). es, l'interprétation se base
Pour le reste des lign

sur la dé t fo essus indiquée u F fini ion des rmules ci-d s comme en ACP o

en A C.

AXES FA 2
CTORIELS 1 A
(TOUTES LES VALEURS SO ULT EE R)
NT M IPLI S PA 1000

CLAS
SE AINE BNJMN ! POIDS INR QLT! F 1 COR CTR!
F 2 COR CTR 32 LOUZ MBIK ! 74 120 892! -197
820 357! 44 41 26 42 37 34! 153 91 783! 29 49 16!
-100 580 284 26 ONDZ LIK1! 75 105 982! 119 346
132! 124 376 212 43 NZAK 41 ! 252 74 974! 42 210
56! 41 191 76 40 36 27 ! 204 55 791! 26 88 18! -32
127 38 44 31 39 ! 241 59 728! -62 537 114! -4 2 1 !
1000 504 ! 693! 637

Tableau 18 : Facteurs pour les six classes de la partition.

En ce qui concerne le tableau 18, il faudra relever les spécificités suivantes : - pour les six classes de la partition, un individu appartient à une classe et une seule ; dans ce cas, la somme des poids des diverses classes est égale au poids total du nuage : 1000/1000 ;

0 q

- par contre, l'inertie relativement à l'origine du centre d'une classe n'est pas la somme des inerties des points constituant la classe ; mais elle est égale à cette somme diminué de l'inertie interne de la classe q . Voilà pourqu INR

oi le total de la colonne est

inférieur à 1 ; ce total représente l'inertie interclasse de la partition retenue (ici ex en primé millième) et l'inertie intra-classes, le comp mentaire à 1 de ce total. De façon analogue la

lé

CTR
somme des sur un facteur donne la part d'inertie interclasse à l'inertie du facteur. ***b)***

Etude des classes par rapport à des variables

Dans cette deuxième analyse, on recherche quelles sont les variables responsables de la distance d'une classe au centre de gravité du nuage. On se place donc ici dans l'espace

q

$CardJ\,j$ correspond un axe, la

des profils sur l'ensemble des variables : à chaque variable

coordonnée sur cet axe étant la composante du profil relative à la variable j . Bien que le nombre de variables ne soient pas élevées, les résultats imprimés sur le listage, occupe une importante surface de papier imprimé. On a donc laissé au programme de ne retenir que les variables ayant les plus fortes contributions aux nœuds supérieurs (tableau 19).

Les coordonnées du centre de gravité du nuage sont celles de la ligne 49, aux colonnes J

$_{IJ}k$

STAS, MATH, GEOE, etc. (profil sur de la ligne de marge du tableau). On peut donc comparer les lignes suivantes : classes 48, 47, 46, et 45 (aux colonnes indiquées) avec la ligne 49 pour savoir en quoi ces classes diffèrent de la classe 49 (centre du nuage). On d

ira par exemple que la classe 46 s'écarte du centre pour un taux moyen en comptabilité d'entreprise et un taux faible en techniques d'expression. On confirme ces résultats en

COR

lisant la colonne . On fini par établir une liste des variables responsables de l'écartement d'une classe au centre du nuage.

Notons que si, toutes les variables avaient été retenues, nous aurions obtenu pour toutes les lignes, . Le fait de n'avoir retenu que quelques variables, cette valeur de la $QLT = 1000$

qualité de représentation est toutes les classes.
descendue au dessous de 1000 et ce, pour

On vérifie que les valeurs des co $POIDS\ INR$

lonnes et sont les mêmes que dans le tableau

17. On peut faire éditer les résultats similair avité des six classes de es pour les centres

de gr

la part n, d nies à partir inq œu les p hau
itio éfi des c n ds lus ts.

(TOUT U ONT L PA 00, ’EX T DE RHO2 ES ES LES VALE RS S MULTIP IEES R 10 A L CEP ION QUI T MULTI E P 10** (

PLI AR 5))

AINE BNJM POIDS INR QLT RHO2 TAS COR CTR! ATH COR CTR

CLASSE 49

46 48 ! 1000 0 0 0! 64 0 0! 58 0 0

! ! S M

48 32 47 ! 673 49 817 212! 66 54 34! 62 115 38 47 42 45 ! 599 47 788 230! 70 227 135! 62 143 46

46 26 43 ! 327 101 817 897! 58 -54 69! 50 -115 79 45 40 44 ! 446 42 783 274! 71 304 160! 58 1 0 ! !

!

! ! 398! 163

CLASSE AINE BNJM ! POIDS INR QLT RHO2 ! GEOE COR CTR! COME COR CTR 49 46 48 ! 1000 0 0 0 ! 58 0 0! 54 0 0 48 32 47 ! 673 49 817 212 ! 58 1 1! 45 -647 216 47 42 45 ! 599 47 788 230 ! 58 6 5! 47 -401 129 46 26 43 ! 327 101 817 897 ! 57 -1 2 ! 72 647 445 45 40 44 ! 446 42 783 274 ! 62 131 99 ! 48 -271 77 ! !

!

! ! 106 ! 86

CLASSE AINE BNJMN ! POIDS INR QLT R TEXP COR CTR ! HO2 !

46 48 ! 1000 0 0 0 ! 51 0 0 ! 49

48 32 47 ! 673 49 817 212 ! 52 1 1 !

47 42 45 ! 599 47 788 230 ! 53 11 11 !

46 26 43 ! 327 101 817 897 ! 51 -1 2 !

45 40 44 ! 446 42 783 274 ! 55 77 65 !

! ! !

! ! 78 !

Tableau 19 : Etude des classes par rapport aux variables initiales (variables ayant les plus fortes contributions aux nœuds supérieurs)

7.2.5.3 Etude des dipôles par rapport à des axes. Formulaire.

$n\ (\ (\ a\ n),b(n))$

A chaque nœud d’une classification, est associé un dipôle formé par les centres de ses deux successeurs immédiats. Dans l’espace rapporté au système d’axes factoriels, on cherche à préciser la situation des segments joignant dans les $a\ n(\ ),b(n)$ dipôles. On calcule :

$$D\ n(\ )\ F\ (a(n)\ b(n))_{\alpha\alpha} = -\ a\ n(\ )$$

- la différence ; elle renseigne sur la position relative de par rapport à .

$b\ n(\ )$

$^{2}{}^{2} COD(\ ) n\ (F\ (a(\ )\ n\ F\ (b\ n(\ ))) /$ [redacted]$_{\alpha\ \alpha}$

= − $_{\alpha}$[redacted] -

$^{2}{}^{2} D\ n(\ )/$ [redacted] = $_{\alpha}$[redacted]

= v n

$^{2}(\ )/\ (\ )\ Q\ D_{n\ n}\ n$

avec : $_{n}$

$_{(\ )(\ )}/\ Q f_{n\ a\ n\ b\ n}$ = ff

$COD(\ )\ n\ _{\alpha}$ α et la droite joignant les centres
est le cosinus carré de l'angle formé par l'axe

des classes et :
$a\ n(\ )\ b\ n(\ )$

$COD(\ )\ n\ _{\alpha} \Rightarrow$ α explique en quasi totalité la séparation entre et ;
si #1 $a\ n(\ )\ b\ n(\ )$

$COD(\ )\ n\ _{\alpha} \Rightarrow$ α n'explique pas la si #0
séparation entre et $a\ n(\ )\ b\ n(\ )$

- $a\ n(\ ) - b(n)$ α rapporté à l'inertie totale sur cet axe (ou l'inertie du dipôle sur l'axe

contribution relative du nœud à l'axe) :

$^{2}(\ )/\ CTD\ Q_{n} D\ n$ = $_{\alpha}\lambda_{\alpha}$

si $CTD(\ )\ n$ #1 $_{\alpha} \Rightarrow$ α la dispersion du nuage sur l'axe α est due exclusivement aux éléments des classes et .
$a\ n(\ )\ b\ n(\ )$

7.2.5.4. Etude des dipôles par rapport à des axes. Exemple

a) Etude des dipôles par rapport à des axes factoriels

On donne dans le tableau 20 les résultats de cette étude. On rappelle que cinq facteurs ont été extraits, deux seulement sont présentés. On constate que, seul le nœud 48 a un COD donc assez nette. De plus, on a sur le plan (1,2) $COD1$+$COD2$ = + 0.679 0.129 = 0.808 (qualité de représentation) : le dipôle est assez proche du plan (1,2).

assez fort (0.679) sur l'axe 1. Sur l'axe 1, la séparation entre $A(48) = 32$ et $B(48) = 47$ est

CTD

L'inertie totale du nuage sur l'axe 1 est expliqué à 29% ((48)=291) par la dichotomie e $A(48) = 32$ $B(48) = 47$ formant la classe 48. En projection sur l'axe ntre les deux classes et 1, la partition de I en deux classes : 32 et 47 a une in · λ_1 ertie interclasse de : 291 (et érifie que le total e la colonne CTD t a 7 λ_1 une iner ie intr -classe de 09 ·). On v d du tableau ne CTR du ableau 8 (asses e la partition). Il en est de 20 est égal à celui de la colon t 1 cl d me r le al de la c ne IND l IND =tota INR mê pou tot olon (tota l =504).

AX CT IEL
ES FA OR S 1 A 2
(TO ES L AL NT PU A)
UT ES V EURS SO LTIPLIES P R 1000
BNJMN ! ! ! COD CTD NŒUD AINE POIDS IND QLD D 1 COD CTD D 2 49 46 48 ! 1000 150 984 !
89 401 217 ! 89 396 318 48 32 47 ! 673 118 875 ! -188 679 291 ! 82 129 82 47 42 45 ! 599 86 580 !
51 116 36 ! -84 319 147 46 26 43 ! 327 79 972 ! 76 148 42 ! 83 176 74 46 40 44 ! 446 72 880 ! 88
409 106 ! -28 40 16 ! ! !
! ! 693 ! 637

Tableau 20 : Etude des dipôles par rapport aux axes factoriels.

b) Etude des dipôles par rapport à des variables ;

Cette étude complète la précédente. Elle permet de déterminer les variables responsables de la séparation des classes. Comme précédemment, on a retenu que les variables ayant les plus fortes contributions aux nœuds supérieurs de la hiérarchie (tableau 21). On signale la présence de la colonne $D\ A2\ B$: c'est le carré de la distance entre les centres de classe $a\ n(\ )$ et $b\ n(\ )$.

rs élevées de la colo COD . Le
Pour l'interprétation, on cherche à repérer les valeu nne par var able $COME$ bilit ; le di ôle dipôle (46,48) est expliqué la i (compta é d'entreprise) p (3) es xp ST (is sa ; le dipôle (26,43) par MATH 2,47 t e liqué par AS Stat tique de la nté)
(mathéma e te de ô n s d'explicat s p icul .
tiqu s) ; le res s dip les 'a pa ion art ières
TO ES L V SONT L L , A L'E PT DE E UT ES ALEURS MU TIP IEES PAR 1000 XCE ION D2AB QUI ST MULTIPLI R
E PA 10**(4)
BNJMN ! POIDS IND QLD D2AB! STAS COD CTD! MATH COD CTD
NŒUD AINE
49 46 48 ! 1000 150 817 198 ! -8 54 102 ! -11 115 117 48 32 47 ! 673 118 341 522 ! -28 231 344 ! -5 10 8
47 42 45 ! 599 86 449 220 ! -6 25 27 ! 16 193 113 46 26 43 ! 327 79 517 395 ! 16 106 105 ! -27 312 167
46 40 44 ! 446 72 264 189 ! -5 22 20 ! -3 10 5 ! ! !
! ! 598 ! 410

NŒUD AINE BNJMN ! POIDS IND QLD D2AB! GEOE COD CTD! COME COD CTD 49 46 48 ! 1000 150
817 198 ! -1 1 3 ! 26 647 661 48 32 47 ! 673 118 341 522 ! -5 7 16 ! -14 71 57 47 42 45 ! 599 86 449 220 !
-14 162 252 ! -3 7 4 46 26 43 ! 327 79 517 395 ! -13 69 97 ! 6 20 10 46 40 44 ! 446 72 264 189 ! 13 164 213
! 4 15 8 ! ! !
! ! 581 ! 740

NŒUD AINE BNJMN! POIDS IND QLD D2AB! TEXP OD CT

C D

49 46 48 ! 1000 150 817 198 ! -1 1 2

48 32 47 ! 673 118 341 522 ! -8 22 53 47 42 45 ! 599 86 449 220 ! -8 61 106 46

26 43 ! 327 79 517 395 ! -5 10 16 46 40 44 ! 446 72 264 189 ! 7 52 76 ! !

! ! 254

Tableau 21 : Etude des dipôles par rapport aux variables initiales.

(**7.2.5.5 Contributions relatives mutuelles entre classes et facteurs**

Notons par $(( ))_{nJ} I N I$ ou par $^{2}( ) M N I_{J}$ l'inertie totale du nuage. On sait déjà que

: $^{2}(( )) (( )) \{ ( ) \}_{nJJ} I N I = = M N I \sum$ v n $n \in N$

$= \sum \{ \}$ λ α $_{\alpha} \in A$

(N ensemble des nœuds et A ensemble des facteurs)

$^{2}(( )) \{ (;) ; \} M N I_{J} = \in \sum$ v αn α A $n \in N$

avec

$^{2}(;) ( )_{n} n Q D n$ v α = $_{\alpha}$

v (; n α) est la contribution absolue mutuelle de n et α .

On peut aussi noter que :

$\{ (n n ;) N \}$ λ v $_{\alpha} = \in \sum$ α

$\sum$

v ( ) n = {v (n;α) α ∈ A}

On ut e es c b ns ves tue en classes et facteurs. C'est pe nfin calculer l ontri utio relati mu

lles tre

le port

rap :

^{2} v (n;α (N (I))

)/ M_{J}

TABLEAU DES CONTRIBUTIONS RELATIVES MUTUELLES SUR LES FACTEURS 1 A 5

NŒUD AINE BNJMN ! Q(N) IND INCUM! F 1 F 2 F 3 F 4 F 5 49 46 48 ! 220 150 150 ! 60

59 20 0 8 48 32 47 ! 66 118 268 ! 80 15 2 0 6 47 42 45 ! 114 86 354 ! 10 27 0 11 2 46 26

43 ! 58 79 433 ! 12 14 34 17 0 46 40 44 ! 111 72 505 ! 29 3 3 22 6

Tableau 22 : Contributions mutuelles : étude des facteurs.

D

ans le tableau 21, on a limité le nombre des variables dans l'étude des dipôles. Cette

é
tude est ici complétée (tableau 23) par le tableau des contributions mutuelles relatives e ntre dipôles et variables. La dernière ligne donne pour chaque variable l'inertie relative de $^{2}( ) M_{J} I$ par rapport à l'axe α .
Les tableaux 22 et 23 n'appellent aucun commentaire particulier.

TABLEAU DES CONTRIBUTIONS MUTUELLES
(TOUTES LES VALEURS SONT MULTIPLIEES PAR 10 **(4))
LA DERNIERE COLONNE DONNE LA PART DE L'INERTIE D'UNE VARIABLE A L'INERTIE TOTALE.

NŒUD AINE BNJMN ! IN(N) STAT MSTA STAS! STAE MATH PROB ECON DEMO 49 46 48 ! 1498 0 81 81 ! 26 172 0 2 56 48 32 47 ! 1182 71 2 273 ! 10 11 5 72 526 47 42 45 ! 860 74 3 21 ! 3 166 91 1 217 46 26 43 ! 785 19 70 83 ! 17 245 8 6 8 46 40 44 ! 719 4 7 16 ! 0 7 18 84 266 ! !
! 336 270 794 ! 192 1469 257 549 1754

NŒUD AINE BNJMN! IN(N) INFO GEOE COME! CON TEXP ANGL HIST STAG 49 46 48 ! 1498 35 2 969 ! 69 1 2 0 2 48 32 47 ! 1182 16 9 84 ! 23 26 14 15 24 47 42 45 ! 860 2 140 6 ! 30 53 2 50 0 46 26 43 ! 785 58 54 15 ! 111 8 40 12 30 46 40 44 ! 719 12 118 11 ! 19 38 2 115 3 ! !
! 297 554 1465 ! 708 498 265 415 176 Tableau 23 : Contributions mutuelles : étude

des variables.

7.2.6. Introduction des nœuds de la classification dans le graphique de l'analyse factorielle.

Une synthèse pratique des procédures factorielles et celles de classification, consiste à situer les classes obtenues par la CAH sur l'espace factoriel. Les coordonnées de ces classes sont les barycentres des éléments qui la composent.

$$n\ ,\ a\ n( )\ ,\ b\ n( )$$

Dans l'espace factoriel on peut représenter, soit les fourches (le triplet issues des classes supérieures, soit encore les classes de la partition retenue (successeurs des classes supérieures). On a choisi ici de ne représenter que les classes supérieures, pour éviter une trop grande densité de points dans l'espace factoriel.

On donne dans le tableau 3.12 les coordonnées des classes dans l'espace factoriel de dimension 5. L'examen de la position de ces classes par rapport aux facteurs permet d'affiner l'interprétation des axes factoriels.

AXES FACTORIELS

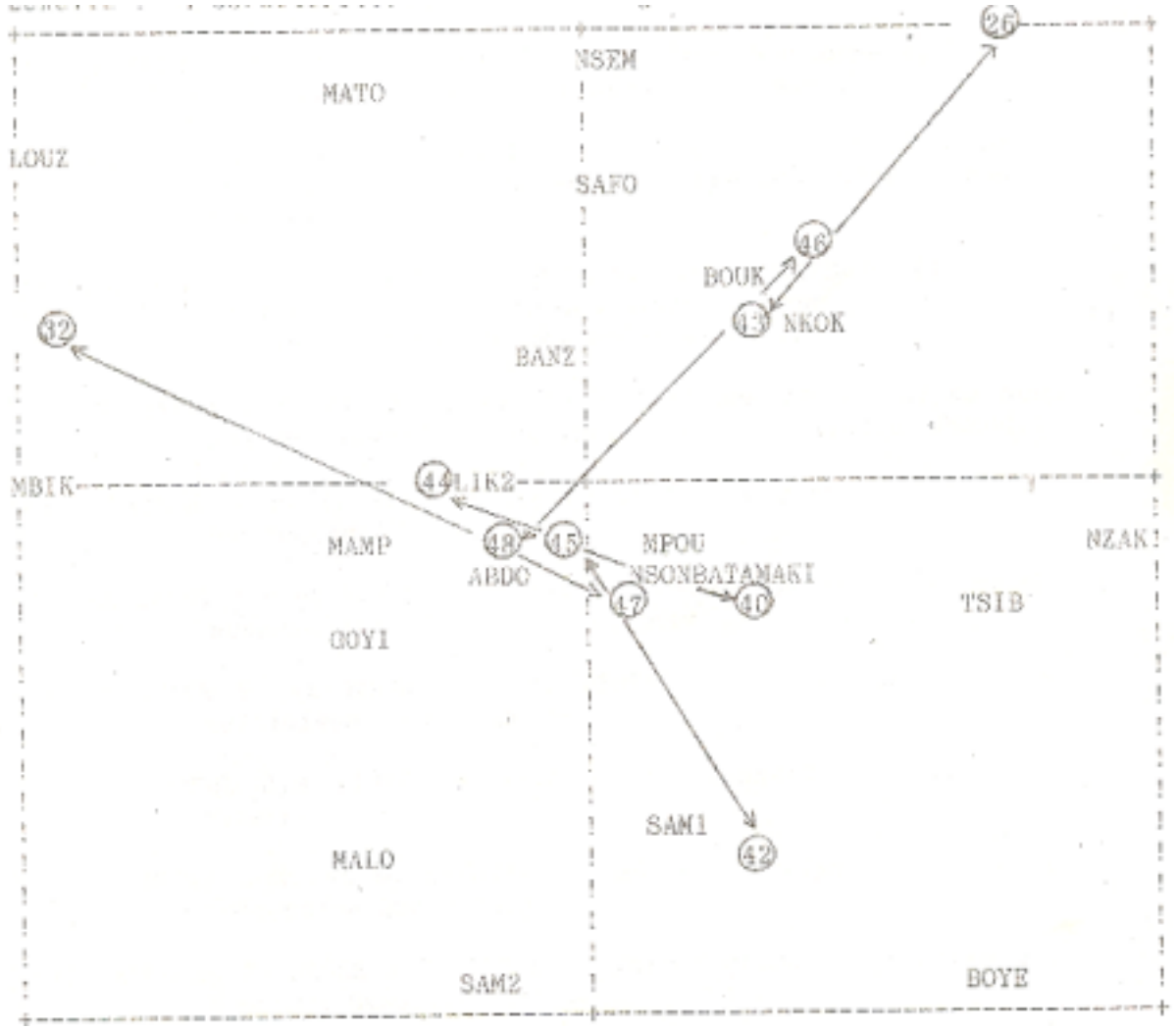
	1	2	3	4	5
N 26	119	124	67	-75	-16
N 27	-35	-30	67	78	-41
N 28	44	-27	8	62	-15
N 29	26	60	-13	7	-50
N 30	-43	-13	-7	-25	64
N 31	-85	-59	-3	-47	-47
N 32	-196	44	-7	12	-34
N 33	16	82	-69	63	35
N 34	-16	-115	-42	-49	47
N 35	-37	0	-2	5	67
N 36	69	-31	13	42	11
N 37	77	-83	79	-22	16
N 38	18	31	-38	-17	-51
N 39	-48	26	16	-5	69
N 40	26	-31	35	56	-10
N 41	17	52	-50	15	-16
N 42	29	-99	16	-36	32
N 43	42	41	-63	16	-22
N 44	-61	-3	9	-19	29
N 45	-20	-16	21	15	11
N 46	60	60	-33	-4	-21
N 47	-7	-37	20	2	16
N 48	-28	-28	17	3	11
N 49	0	0	0	0	0

Tableau 24 : Coordonnées des classes dans l'espace factoriel x 1000

Le graphique 10 donne la représentation de l'ensemble des 25 étudiants, des cinq fourches principales extraites de la classification, dans l'espace factoriel (1,2). On a relié par un segment de droite les nœuds que la classification a agrégés ensemble. Bien que tous les nœuds ne soient pas représentés, on retrouve les mêmes oppositions rencontrées lors de l'analyse factorielle des correspondances.

AXE HORIZONTAL (1) – AXE VERTICAL (2). Nombre de points : 35

Echelle : 4 caractère(s)= .021 1 ligne = .009



Nombre de points superposés : 3

LIK1(26) ONDZ(26) NGUI(BATA)

Graphique 10 : Représentation dans l'espace factoriel (1,2) des fourches issues des classes supérieures.

Bibliographie

[1] **Benzécri, J.-P., & Coll**., (1980). *L'analyse des données,* Tome 1 *: la taxinomie,* Dunod

[2] **Benzécri, J.-P., & Coll.**, (1980). *L'analyse des données,* Tome 2 *: l'analyse des correspondances,* Dunod

[3] **Benzécri, J.-P., Benzécri F.,** (1980). *Pratique de l'analyse des données,* Tome 1 *: Analyse des correspondances. Exposé élémentaire,* Dunod

[4] **Benzécri, J.P., Bastin, CH., Bougarit, CH., Cazes, P.,** (1980). *Pratique de l'analyse des données,* Tome 2 *: Abrégé théorique. Etudes de cas modèles,* Dunod

[5] **Benzécri, J.-P., & Coll.**, (1981). *Pratique de l'analyse des données,* Tome 3 *: Linguistique et lexicologie,* Dunod

[6] **Benzécri, J.-P.**, (1982). *Histoire et préhistoire de l'analyse des données,* Dunod [7] **Benzécri, J.-P., & Coll.**, (1986). *Pratique de l'analyse des données,* Tome 5 *: Economie,* Dunod

[8] **Bertier, P., Bouroche, J.M.,** (1975)- *Analyse des données multidimensionnelle,* PUF

[9] **Bouroche, J.M.,** (1977). A*nalyse des données en marketing,* Masson [10] **Bouroche, J.M., Saporta, G.,** (1980). *l'analyse des données,* Collection Que sais-je ? PUF.

[11] **Cailliez, F. Pages J.P.,** (1976). *Introduction à l'analyse des données, Smash.* [12] **Cazes, P., Lecoutre, J.P.,** (1977). Etude de quelques problèmes de codage en analyse des correspondances, *Cahiers du Bureau universitaire de recherche opérationnelle*, n°27 pp.49-66.

[13] **Cazes, P.,** (1980). L'analyse de certains tableaux rectangulaires décomposés en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. II Questionnaire : variantes de codages et nouveaux calculs de contributions, *Cahiers de*

l'Analyse des données, Vol 5 n°4 pp. 387-403.

[14] **Cazes, P.,** (1982). Note sur les éléments supplémentaires en analyse des correspondances : I Pratique et utilisation, *Cahiers de l'Analyse des données*, Vol 7 n°1 pp. 9-23. II Tableaux multiples, *Cahiers de l'Analyse des données*, Vol 7 n°2, pp.133-154. [15] **Cazes, P.,** (1983). L'analyse des correspondances multiples. Application à l'étude des questionnaires, *Bulletin de l'ADDAD* n°12.

[16] **Cehessat, R.,** (1976). *Exercices commentés de statistique et informatique appliquée*, Dunod

[17] **Cibois, PH.,** (1987). *L'analyse factorielle*, Collection Que sais-je ? Puf [18] **Celeux, G., Diday, E ; et Ali.,** (1989). *Classification automatique des données*. Environnement statistique et informatique, Dunod.

[19] **Chandon, J.L., Pinson, S.,** (1981). *Analyse typologique. Théories et applications*. Masson

[20] **CNRS.,** (1955). *L'analyse factorielle et ses applications*.

[21] **Corroyer, D.,** (1991). *DS3.* Un logiciel pour le traitement informatique et statistique des données et son enseignement. Apetisd (68, av. de la Faisandrie. 91800 Brunoy). [22] **Corroyer, D., Pierre-Puyesegur, M.A.,** (1992). *L'analyse statistique et informatique des tableaux de contingence*, Apetisd

[23] **Dervin, C.,** (1990). *Comment interpréter les résultats d'une analyse factorielle des correspondances.* ITCF.

[24] **Diday, E., Lemaire, J., Pouget, J., Testu, F.,** (1982). *Eléments d'analyse de données,* Dunod

[25] **Droesbeke, J-J., Tassi, PH.,** (1990). *Histoire de la Statistique.* Collection Que sais-je ? Puf

[26] **Ducimetiere, P.,** (1970). Les méthodes de la classification numérique. *Revue de Statistique appliquée*. Vol XVIII n°4, pp.5-25.

[27] **Escofier-Cordier, B**., (1965). L'analyse factorielle des correspondances. *Cahiers du Bureau universitaire de recherche opérationnelle,* n°13.

[28] **Escofier, B., Pages, J.,** (1988) – *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation,* Dunod.

[29] **Escoufier, Y.,** (1978). *Cours d'analyse des données,* Montpellier.

[30] **Evrard, Y., Lemaire, P.,** (1976) – *Information et décision en marketing,* Dalloz [31] **Fenelon, J.P.,** (1981). *Qu'est-ce que l'analyse des données ?,* Lefonen. [32] **Foucart, T.,** (1981). *Analyse factorielle, programmation sur micro-ordinateurs,* Masson [33] **Foucart, T.,**

(1984). *Analyse factorielle de tableaux multiples,* Masson [34] **Gendre, F.,** (1976). *L'analyse statistique multivariée,* Librairie Droz. [35] **Gibrat, R.,** (1978). L''analyse des données : Première partie : *Journal de la Société de Statistique de Paris* n°3, pp.201-228. Deuxième partie : les sciences humaines : impasse, échecs et succès. *Journal de la Société de statistique de Paris* n°4 pp.312-331. [36] **Grelet, Y., Lebeaux, M.O.,** (1980). Lecture commentée de sorties de programmes en analyse des données, *Bulletin de l'ADDAD* n°11.

[37] **Jambu, M.,** (1978). *Classification automatique pour l'analyse des données,* Tome 1 : Méthodes et algorithmes, Dunod.

[38] **Jambu, M., Lebeaux, M.O.,** (1978). *Classification automatique pour l'analyse des données,* Tome 2 : Logiciels

[39] **Jambu, M.,** (1989). *Exploration informatique et statistique des données,* Dunod. [40] **LADDAD.,** *Logiciel de l'Association pour le Développement et la Diffusion de l'Analyse des données.* (ADDAD, 22 rue Charcot, Paris 75013)

[41] **Lebart, L., Fenelon, J.P.,** (1971). *Statistique et informatique appliquée,* Dunod

[42] **Lebart, L.,** (1975). *Validité des résultats en analyse des données,* Rapport CREDOC-DGRST.

[43] **Lebart, L., Morineau, A., Tabard, N.,** (1977). *Techniques de la description statistique.* Méthodes et logiciels pour l'analyse de grands tableaux, Dunod

[44] **Lebart, L., Morineau, A., Fenelon, J.P.,** (1979). *Traitement des données statistiques.* Dunod

[45] **Lebart, L., Salem, A.,** (1988). *Analyse statistique des données textuelles,* Dunod.

[46] **Lefebvre, J.,** (1980). *Introduction aux analyses statistiques multidimensionnelles,* Masson

[47] **Lerman, I.C.,** (1981). *Classification et analyse ordinale des données,* Dunod.

[48] **Masson, M.,** (1980). *Méthodologies générales de traitement statistique de l'information de masse,* Cedic/Nathan.

[49] **Morineau, A.,** (1983). *Lecture commentée d'une analyse de correspondances multiples suivie d'une classification (Programme SPAD),* Cisia.

[50] **Morlat, G.,** (1976) – *Préface de l'introduction à l'analyse des données.* Smash [51] **Moscarola, J.,** (1990) – *Enquêtes et analyse de données.* Vuibert

[52] **Nakache, J.P., Chevalier, A., Morice, V.,** (1981). *Exercices commentés de mathématiques pour l'analyse statistique des données*, Dunod.

[53] **Pages, J.P., Cailliez, F., Escoufier, Y.,** (1979). Analyse factorielle : un peu d'histoire

et de géométrie. *Revue de Statistique Appliquée,* Vol XXVII, n°1 pp. 5-28. [54] **Philippeau, G.,** (1986). *Comment interpréter les résultats d'une analyse en composantes principales.* ITCF.

[55] **Pontier, J., Dufour, A.B., Normand, M.,** (1990). *Le modèle euclidien en analyse des données,* Ellipses.

[56] **Robert, C.,** (1989). *Analyse descriptive multivariée.* Application à l'intelligence artificielle, Flammarion.

[57] **Saporta, G.,** (1990). *Probabilités Analyse des Données et Statistiques.* Editions Technip.

[58] **STATPC.,** (1989). Logiciel de traitement statistique : méthodes graphiques et numériques (Bleuse-Trillon B. 10, rue Croix de Malte 45000 Orléans). [59] **Torrens – Ibern, J.,** (1972). *Modèles et méthodes de l'analyse factorielle,* Dunod

[60] **Volle, M.,** (1981). *Analyse des données.* Economica.