

**BUREAU D'APPLICATION DES METHODES
STATISTIQUES ET INFORMATIQUES**

BAMSI REPRINT 04/2003

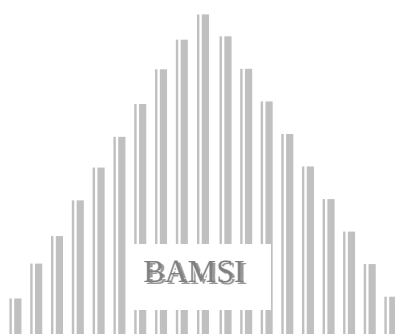
Introduction à l'analyse des données

Samuel AMBAPOUR

**D
o
c
u
m
e
n
t

d
e

T
r
a
v
a
i
l**



BAMSI B.P. 13734 Brazzaville

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

1. INTRODUCTION

2. UN PEU D'HISTOIRE

3. TYPES DE TABLEAUX ANALYSABLES

4. ANALYSE GENERAL

5. L'ANALYSE EN COMPOSANTES PRINCIPALES

5.1. Les données – Les objectifs

5.2. La méthode

5.2.1. Le tableau de données

5.2.2. Analyse des points individus i de $N_j(I)$ dans R^p

5.2.3. Analyse des points individus j de $N_i(J)$ dans R^n

5.2.4. Relation entre les points i de $N_j(I)$ et j de $N_i(J)$

5.2.5. Analyse des points supplémentaires

5.3. Interprétation de l'Analyse en Composantes Principales

5.3.1. Tableau des données de base

5.3.2. Matrice de corrélations des variables

5.3.3. Vecteurs et valeurs propres de la matrice de corrélation

5.3.4. Tableau des facteurs sur I

5.3.5. Tableau des facteurs sur J

5.3.6. Représentations graphiques

6. L'ANALYSE FACTORIELLE DES CORRESPONDANCES

6.1. Les données – Les objectifs

6.2. La méthode

6.2.1. Le tableau de données

6.2.2. Analyse des points i de $N_j(I)$ dans R^p

6.2.3. Analyse des points j de $N_i(J)$ dans R^n

6.2.4. Relations entre les points i de $N_j(I)$ et les points j de $N_i(J)$

6.2.5. Eléments supplémentaires

6.3. Interprétation d'une analyse factorielle des correspondances

6.3.1. Tableau des données de base

6.3.2. Vecteurs et valeurs propres

6.3.3. Tableaux des facteurs sur I et sur J : aides à l'interprétation

6.3.4. Représentations graphiques

6.4. Analyse des correspondances multiples

6.4.1. Tableau disjonctif complet

6.4.2. Tableau de Burt

6.4.3. Equivalence entre les deux analyses précédentes

6.4.4. Calcul de contributions dans le tableau disjonctif complet

6.4.5. Interprétation d'une analyse des correspondances multiples

6.4.5.1. Tableau des données de base

6.4.5.2. Valeurs propres

6.4.5.3. Tableaux des facteurs sur i et J

6.4.5.4. Représentation graphique

7. CLASSIFICATION ASCENDANTE HIERARCHIQUE

7.1. Principes généraux

7.1.1. Partition et hiérarchie

7.1.2. Classification ascendante et classification descendante

7.1.3. Construction d'une classification ascendante hiérarchique

7.1.4. Critères d'agrégation

7.2. L'interprétation d'une classification ascendante hiérarchique

7.2.1. Le tableau des données

7.2.2. Histogramme des indices de niveau de la hiérarchie

7.2.3. Le tableau du contenu des classes

7.2.4. Représentation de la classification ascendante hiérarchique

7.2.5. Calcul de contributions

7.2.5.1. Etude des classes par rapport à des axes. Formulaire

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

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

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

7.2.5.5. Contributions relatives mutuelles entre classes et

facteurs

7.2.6. Introduction des nœuds de la classification dans le graphique

de l'analyse factorielle

REFERENCES BIBLIOGRAPHIQUES

“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 dégagant 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 en 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. Benzécri : "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 ordinaire à l’aide des variables nominales).

Ces méthodes ont trouvé leurs applications surtout dans le domaine du marketing[9].

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 jugements de préférence globale sur un ensemble J 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 -ème 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 i sur toutes les régions, relativement à l'effectif de la classe d'âge j . Dans le tableau de préférence rencontré ci-haut, la ligne 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 différents"[31].

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

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 l'ensemble d'industries (ou produits), J un ensemble de pays, T un ensemble d'époques, k_{IJT} désignant les échanges pour le produit i , à l'instant t en provenance (ou à destination) du pays j . Une généralisation au cas quaternaire a été étudiée et on obtient un tableau de la forme k_{IJPT} où I est par exemple l'ensemble des pays exportateurs, J l'ensemble des mêmes pays considérés comme exportateurs, P est un ensemble des classes de produits et T un ensemble d'époques : k_{IJPT} est donc la valeur des importations du pays i en provenance du pays j (ou des exportations du pays j à destination du pays i), rentrant dans la classe de produits p et effectuées en l'année t . Pour 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 J . On a $CardI$ observations sur lesquelles sont mesurées $CardJ$ variables : x_{ij} est la mesure de la variable j de J sur l'individu i de I . ($CardI = n, CardJ = p$) . x_{ij} peut être la note 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 p points correspondant chacun à 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 n 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 F_1 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 :

$$X'Xu_\alpha = \lambda_\alpha u_\alpha \text{ dans } R^p$$

Le vecteur u est norme par la relation :

$$u'u = 1$$

Le premier axe factoriel est donc le vecteur u_1 correspondant λ_1 la plus grande valeur propre de $X'X$. 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 q premiers vecteurs propres $u_1, \dots, u_q$ de $X'X$. 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 $\lambda_\alpha \neq 0$,
 $u_\alpha = \lambda_\alpha^{-1/2} X'v_\alpha$ est vecteur unitaire de $X'X$ relatif à la même valeur propre.

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

u_α est appelé α -ème *axe factoriel* dans R^p .

v_α est appelé α -ème *axe factoriel* dans R^n .

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 p variables 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.

$$X = \{x_{ij} | i \in I, j \in J\}$$

	x_j
	.
	.
	.
i	 x_{ij}
	.
	.
	.

On calcule :

i)- La moyenne de la variable x_j :

$$\bar{x}_j = \sum \left\{ \frac{m_i}{M} x_{ij} | i \in I \right\} \text{ avec } M = \sum \{m_i | i \in I\}$$

où m_i est le poids affecté à l'individu i ; $m_i = \frac{1}{\text{Card}I}$ et $\sum \{m_i | i \in I\} = M = 1$

ii)- La variance de la variable x_j :

$$\sigma_j^2 = \text{var}(x_j) = \sum \left\{ \frac{m_i}{M} (x_{ij} - \bar{x}_j)^2 | i \in I \right\}$$

iii) La variable centrée et réduite qui a pour composantes sur l'ensemble I :

$$X_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

où σ_j est l'écart type de x_j

$$\Rightarrow \text{moy}(X_j) = \sum \{X_{ij} | i \in I\} = 0$$

$$\text{et } \text{var}(X_j) = \sum \left\{ [X_{ij} - \text{moy}(X_j)]^2 | i \in I \right\} = 1$$

iv)- Le coefficient de corrélation linéaire entre deux variables x_j et $x_{j'}$:

$$r_{jj'} = \sum \left\{ \frac{m_i}{M} X_{ij} X_{ij'} | i \in I \right\}$$

qui prend les valeurs entre -1 et $+1$.

5.2.2. Analyse des points individus i de $N_j(I)$ dans R^p

On se placera au centre de gravité du nuage des points de base (normalisation centrée-réduite). Le i -ème 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) m_i . Si l'on note par :

$$N_j(I) = \left\{ (X_{ij}; \frac{m_i}{M}) | i \in I \right\} \text{ le nuage des points } i \in I ;$$

i) Le centre de gravité G de ce nuage a pour j -ème coordonnée :

$$\begin{aligned} X_{Gj} &= \sum \left\{ \frac{m_i}{M} X_{ij} | i \in I \right\} \\ &= \frac{1}{\sigma_j} \left[\sum \left\{ \frac{m_i}{M} x_{ij} | i \in I \right\} - \bar{x}_j \sum \left\{ \frac{m_i}{M} | i \in I \right\} \right] = 0 ; \end{aligned}$$

c'est donc l'origine du système d'axes dans lequel est placé le nuage des individus.

ii) La distance entre deux points de $N_j(I)$ s'écrit dans R^p :

$$\begin{aligned} d^2(i, i') &= \sum \left\{ (X_{ij} - X_{i'j})^2 | j \in J \right\} \\ &= \sum \left\{ \left(\frac{x_{ij} - x_{i'j}}{\sigma_j} \right)^2 | j \in J \right\} \end{aligned}$$

(c'est la distance euclidienne "usuelle"). Ainsi chaque variable aura une contribution égale à la dispersion totale du nuage $N_j(I)$.

iii) La distance d'un point de i au centre de gravité G du nuage $N_j(I)$ vaut :

$$d^2(i, G) = \rho^2(i) = \sum \{X_{ij}^2 | j \in J\}$$

iv) L'inertie d'un point i par rapport au centre de gravité est :

$$I_n(I) = \frac{m_i}{M} \rho^2(i)$$

et l'inertie du nuage $N_J(I)$ sera égal à :

$$\begin{aligned} I_n(N_J(I)) &= \sum \left\{ \frac{m_i}{M} \rho^2(i) | i \in I \right\} \\ &= \sum \left\{ \frac{m_i}{M} (\sum_{j \in J} X_{ij}^2) | i \in I \right\} \\ &= \sum \sum \left\{ \frac{m_i}{M} X_{ij}^2 | i \in I, j \in J \right\} \\ &= \sum \{ \text{Var}(X_j) | j \in J \} \end{aligned}$$

or $\text{var}(X_j) = 1$

$$\Rightarrow I_n(N_J(I)) = \text{Card}J$$

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 $r_{jj'}$. C'est donc cette matrice qu'il faudra diagonaliser pour la recherche des vecteurs et valeurs propres.

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

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

Les facteurs sont de moyenne nulle, de variance égale à λ_α , et sont deux à deux orthogonaux.

En effet :

$$\sum \left\{ \frac{m_i}{M} F_\alpha(i) | i \in I \right\} = 0$$

$$\lambda_\alpha = \sum \left\{ \frac{m_i}{M} F_\alpha^2(i) | i \in I \right\}$$

et

$$\sum \left\{ \frac{m_i}{M} F_\alpha(i) F_\beta(i) | i \in I \right\} = 0 \quad \text{si } \alpha \neq \beta$$

On sait déjà que la somme de toutes les valeurs propres est égale au nombre $\text{Card}J$ de variables. Et donc :

$$I_n(N_J(I)) = \sum \lambda_\alpha = \text{Card}J$$

5.2.3. Analyse des points variables j de $N_I(J)$ dans R^n

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 nuage des variables. On a vu que $Var(X_j)=1$ c'est-à-dire que $d^2(0, j)=1$; 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 $N_I(J)$ dans R^n :

$$d^2(j, j') = \sum \{(X_{ij} - X_{ij'})^2 | i \in I\}$$

En tenant compte du fait que $Var(X_j)=Var(X_{j'})=1$ et $\sum X_{ij}X_{ij'} = r_{jj'}$, On trouve que :

$$d^2(j, j') = 2(1 - r_{jj'})$$

où $r_{jj'}$, est le coefficient de corrélation linéaire entre les variables j et j' . Ainsi, les proximités entre points variables s'expriment en termes de corrélations :

$r_{jj'} = 1 \Rightarrow$ les points j et j' sont confondus ;

$r_{jj'} = -1 \Rightarrow$ les points j et j' sont diamétralement opposés sur la sphère (0,1) ;

$r_{jj'} = 0 \Rightarrow$ les points j et j' sont orthogonaux et se trouvent aux extrémités d'un arc de 90° .

5.2.4- Relation entre les points de $N_J(I)$ et j de $N_I(J)$

Nous avons vu au chapitre 4 les relations qui existent entre les matrices $X'X$ et XX' en ce qui concerne les vecteurs et les valeurs propres. En utilisant ces propriétés, on peut établir les relations de transition entre les facteurs $F_\alpha(i)$ de I et $G_\alpha(j)$ de J . On a :

$$F_\alpha(i) = \lambda^{-1/2} \sum X_{ij} G_\alpha(j)$$

et

$$G_\alpha(j) = \lambda^{-1/2} \sum X_{ij} F_\alpha(i)$$

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 données pour représenter [14] :

- soit une observation relevée dans des conditions douteuses (ou différentes des autres 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 ;

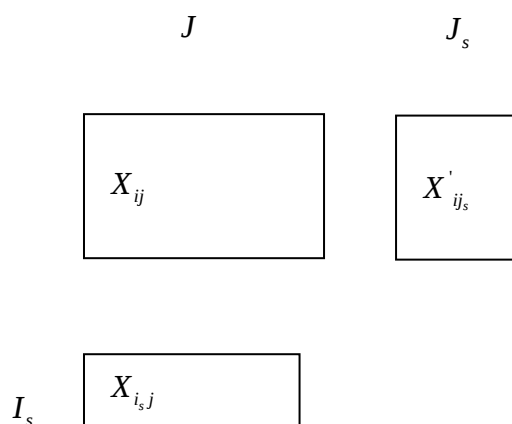
- soit un cas nouveau ;
- soit des éléments de nature différente de ceux analysés.

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-démographiques en supplémentaires.

Considérons la figure suivante :



Si l'on effectue l'analyse en composantes principales du tableau X_{ij} (tableau principal), on peut projeter sur les axes factoriels ainsi trouvés les ensembles I_s (ensemble des individus supplémentaires) et J_s (ensemble des variables supplémentaires). Les coordonnées des individus supplémentaires $i_s \in I_s$ sont les composantes du vecteur $(X_{i_s j})'_\alpha$ et, celles des variables supplémentaires $j_s \in J_s$ les composantes du vecteur $(X_{ij_s})'_\alpha$ (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

Traitions 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 x_{ij} 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.

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

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 *CardJ* 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 :

$$(\lambda_1 - \lambda_2) / \lambda_1 = (3.07452 - 2.66202) / 3.07452 = 0.13417$$

$$(\lambda_2 - \lambda_3) / \lambda_2 = (2.66202 - 2.32011) / 2.66202 = 0.12844$$

$$(\lambda_3 - \lambda_4) / \lambda_3 = (2.32011 - 1.96808) / 2.32011 = 0.15173$$

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 $\alpha+1$. 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].

! 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.

$m_i = 1$; $M = \sum m_i = 25$; $p_i = \frac{m_i}{M} = \frac{1}{25} = 0,04$ (ici 40 exprimé en millièmes).

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

$$I_n(i) = p_i \rho^2(i)$$

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

Ensuite on trouve pour chaque facteur :

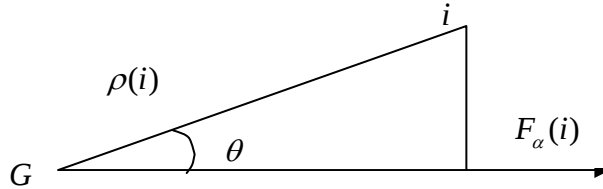
iv) $\#F(F_\alpha(i))$, coordonnées des individus ; l'examen de ces coordonnées permet de connaître comment se répartissent les individus, ceux qui interviennent sur l'axe du côté positif ou du côté négatif.

v) *CTR*, contribution relative de l'individu i et à l'inertie expliquée par l'axe α .

$$CTR(i, \alpha) = p_i F_\alpha^2(i) / \lambda_\alpha ;$$

on remarque que *CTR* varie comme $F_\alpha^2(i)$: les points les plus contributifs sont les plus 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 α . *COR* peut être interprétée comme le cosinus de l'angle formé par un point avec sa projection sur le plan.



$$\cos^2 \theta = \frac{F_\alpha^2(i)}{\rho^2(i)}$$

En prenant toujours le cas de l'étudiant LOUZ on voit que $\cos^2(LOUZ, F_1) = 453$; si on appelle 1000 la situation de LOUZ, on en trouve 453 sur le facteur 1. On peut vérifier facilement que : $QLT = \sum_{\alpha=1}^p COR_\alpha(i) = 1000$ (en millièmes). Comme on le constate les *COR* s'additionnent en ligne ; sommés sur les 17 facteurs, on trouverait 1000 ; sommés sur les 5 (on a extrait 5 facteurs, seulement trois sont imprimés) facteurs, on trouve $QLT = 885$ pour LOUZ..

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 $COR(\alpha, j) = G_\alpha^2(j)$. On interprètera donc l'axe en fonction des variables qui lui sont corrélées. Comme le nuage $N_I(j)$ est situé dans une sphère de rayon 1, l'usage des CTR 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 projection.

!	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 $N_j(I)$ et de $N_j(J)$ dans un espace de faible dimension par rapport aux dimensions d'origine. Ces représentations se font dans la plupart des cas dans un espace à deux dimensions :

5.3.6.1. Représentation graphique associée aux points i

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[10]. L'attention sera alors attirée par l'allure générale de la répartition de l'ensemble des individus.

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).

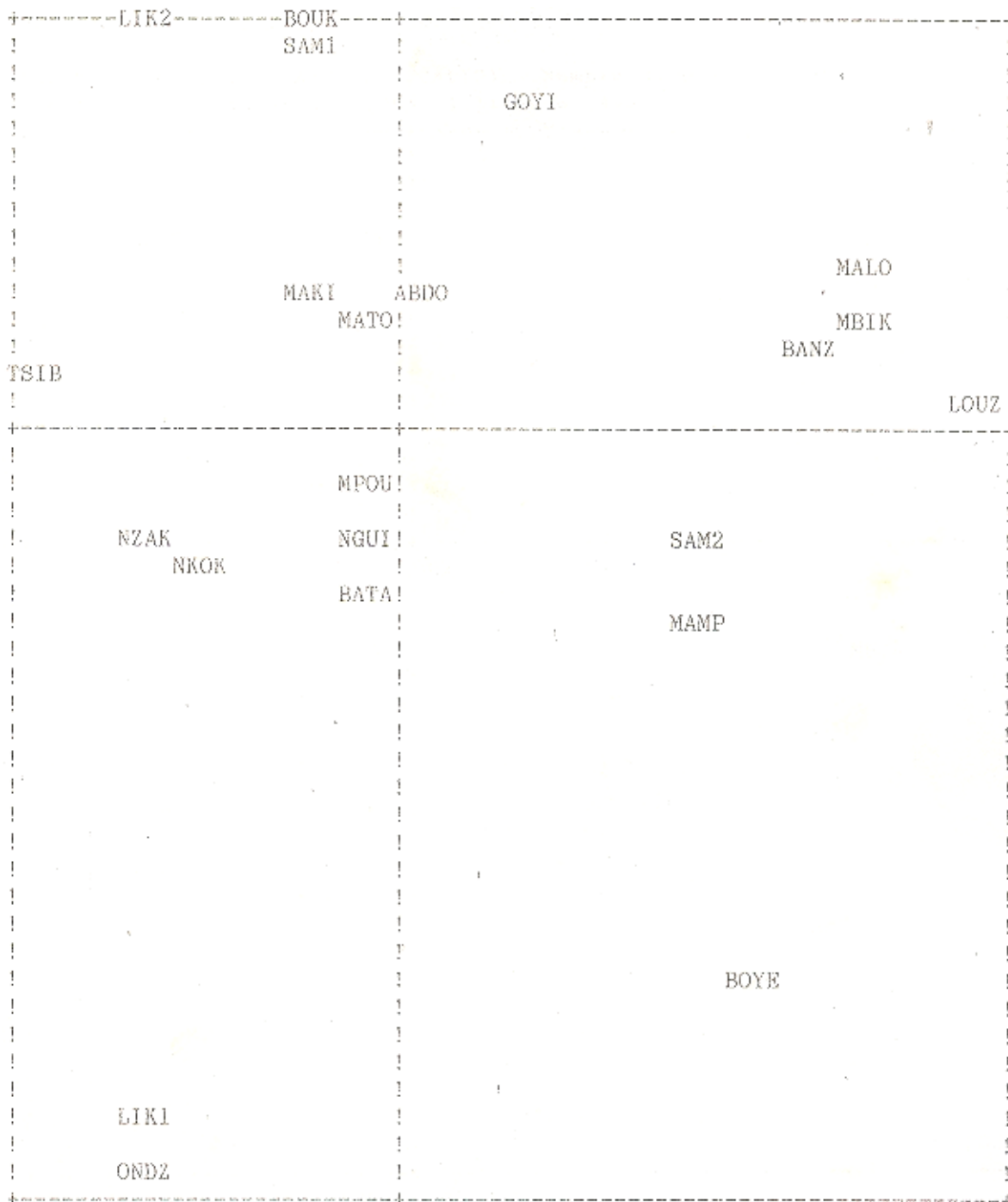
Le troisième facteur peut être considéré comme facteur de culture générale. Sont effectivement corrélées avec ce facteur les variables histoire (HIST), anglais (ANGL), techniques d'expression (TEXP) et la géographie (GEOE).

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



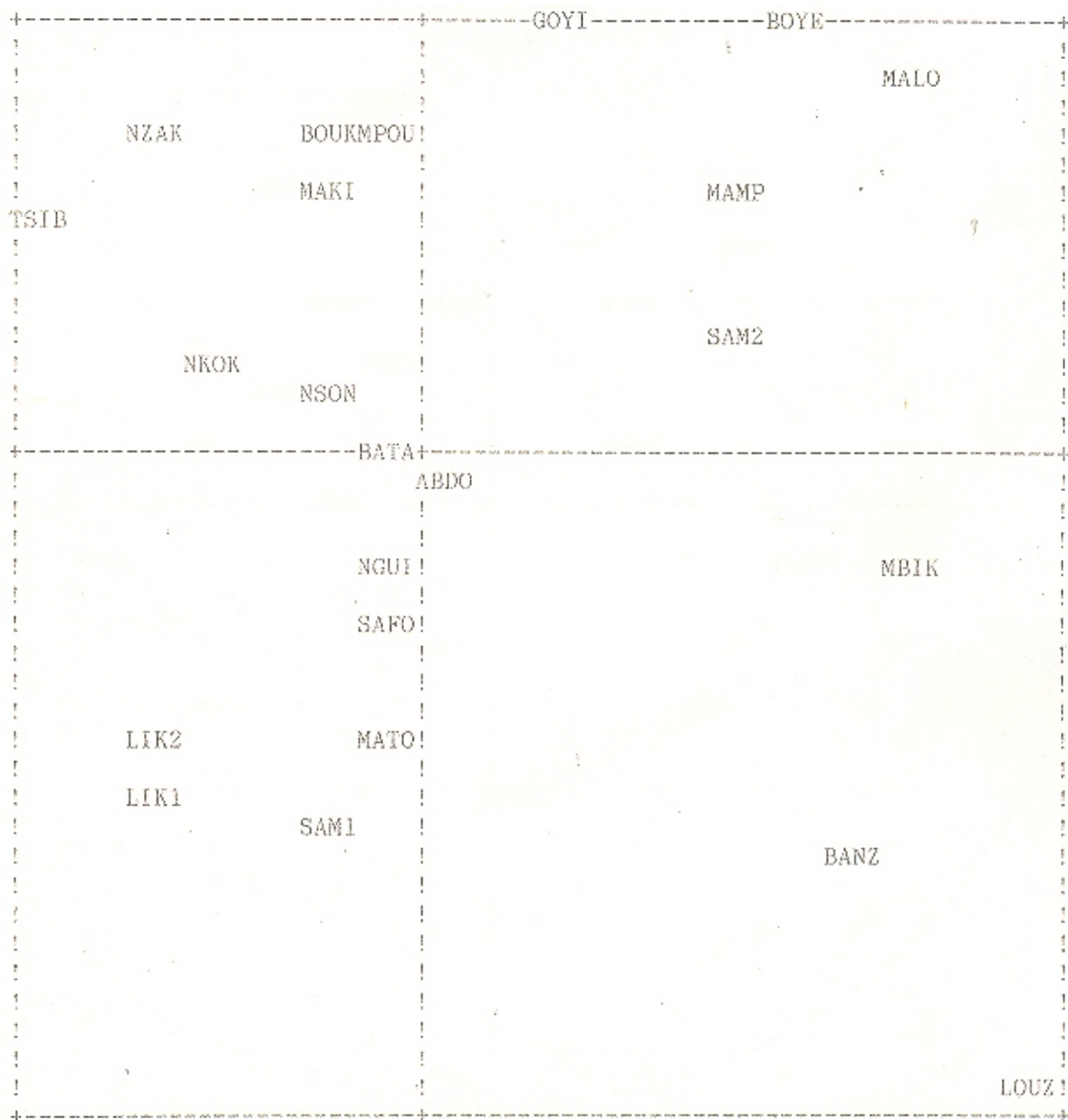
Nombre de points superposés : 3

NSEM(LIK2) NSON(MAKI) SAFO(MPOU)

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

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

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



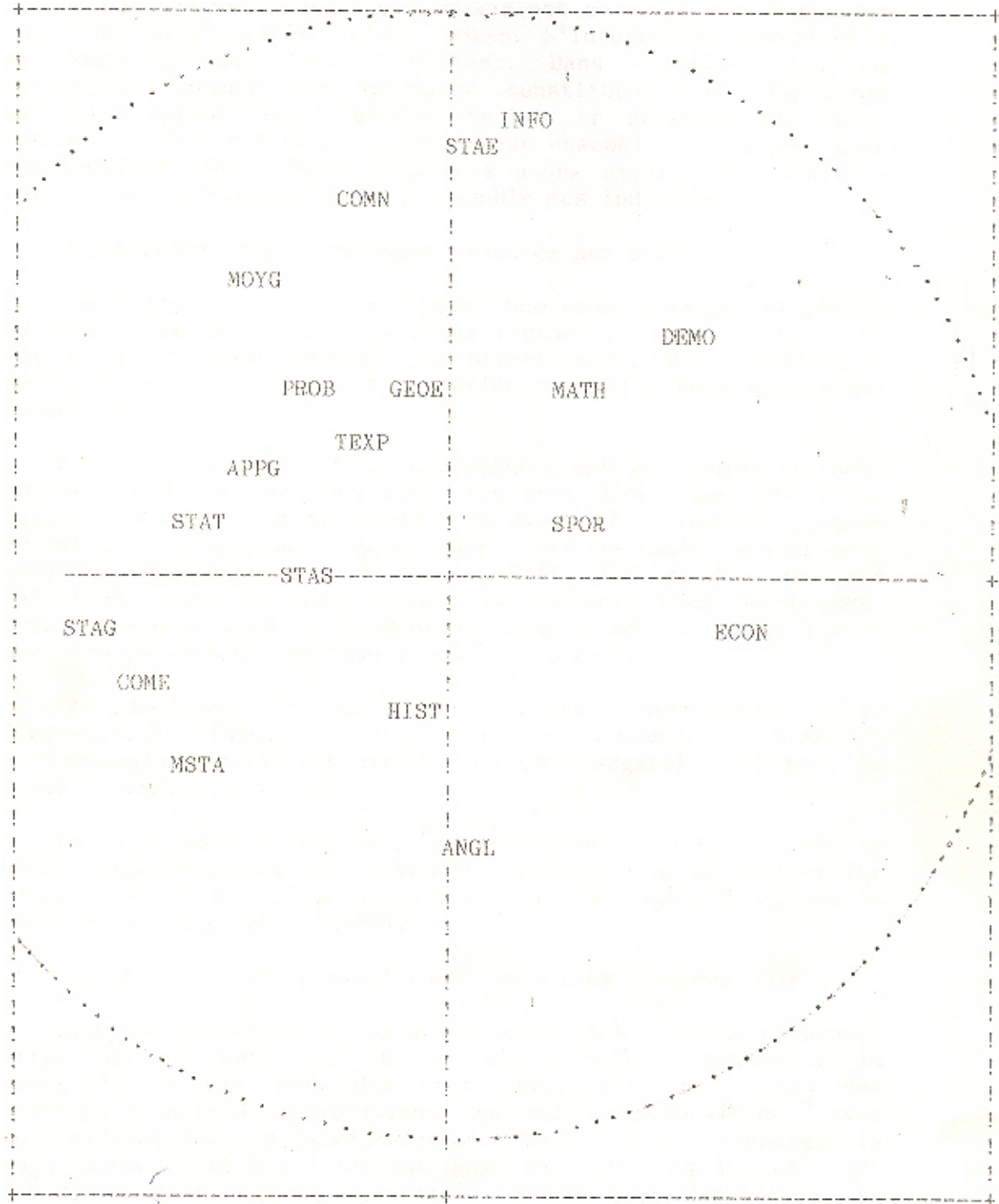
Nombre de points superposés : 2

NSEM(LIK2) ONDZ(LIK2)

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

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

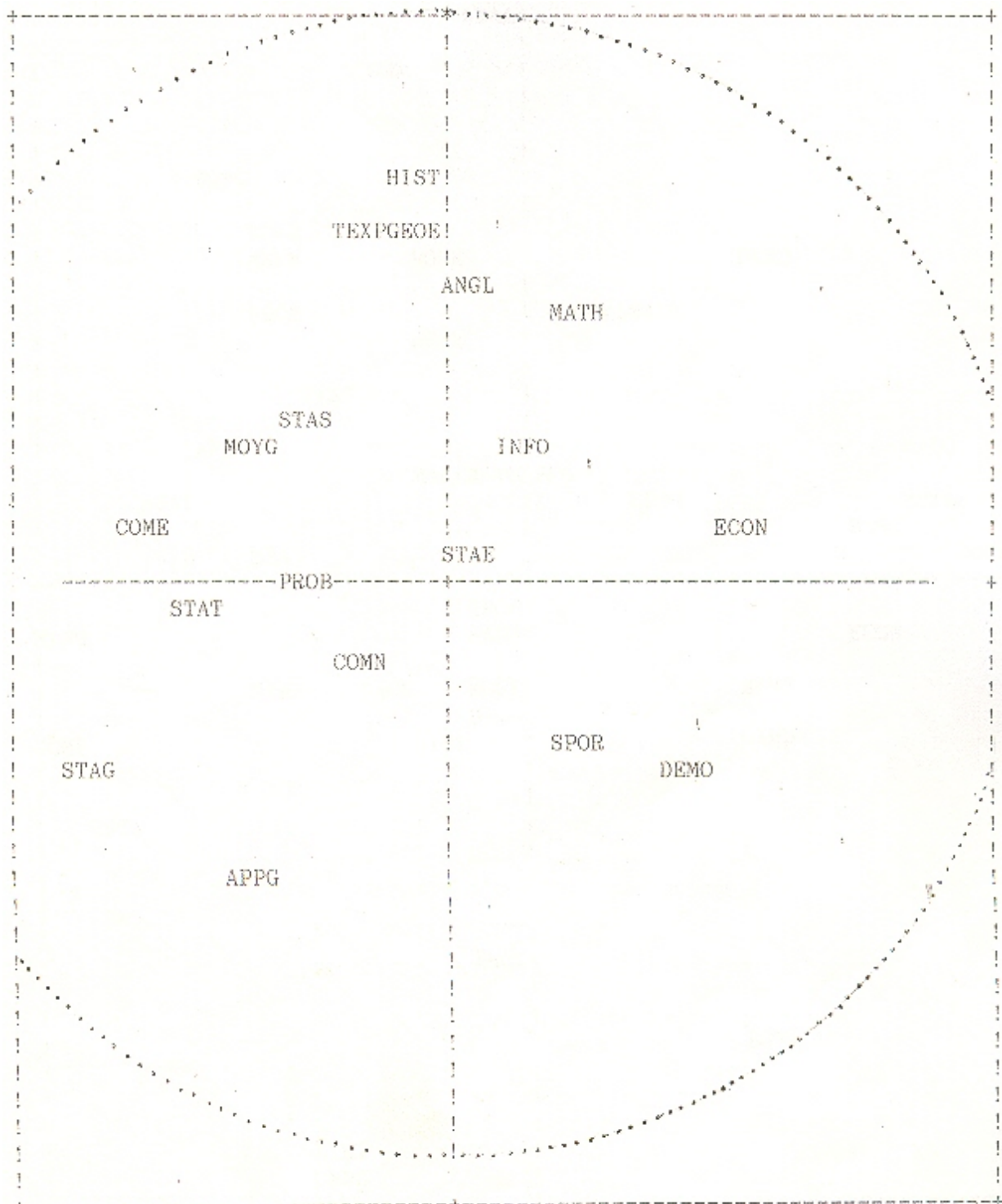
Echelle : 4 caractères = .111 1 ligne = .046



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



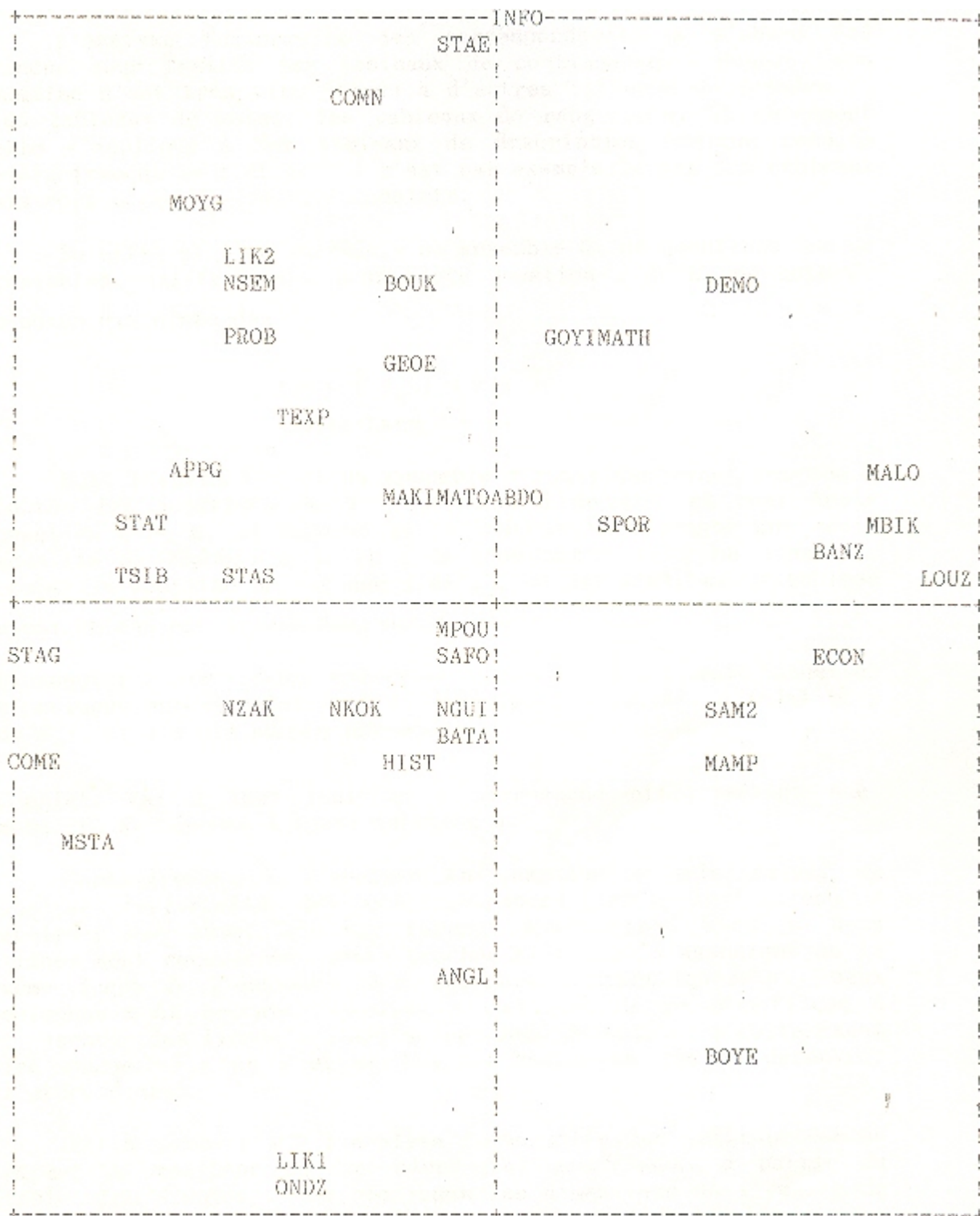
Nombre de points superposes: 1

MSTA (STAT)

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

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 q de Q , on note J l'union disjointe des J_q :

$$J = \bigcup \{J_q | q \in Q\} \text{ (avec } CardJ = p \text{)}.$$

Soit I ($CardI = n$) un ensemble d'individus ayant répondu à toutes les questions de Q . Pour tout i et de I , et pour toute question q de Q , on suppose que l'individu i a adopté une seule modalité de réponse à q , et l'on code par 1 si l'individu i a choisi la modalité de réponse J de J_q , 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 J_q , il n'a pas adopté une modalité j' ($j' \neq j$) de J_q .

Complet, car à tout individu i correspond effectivement une modalité de réponse à toute 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 :

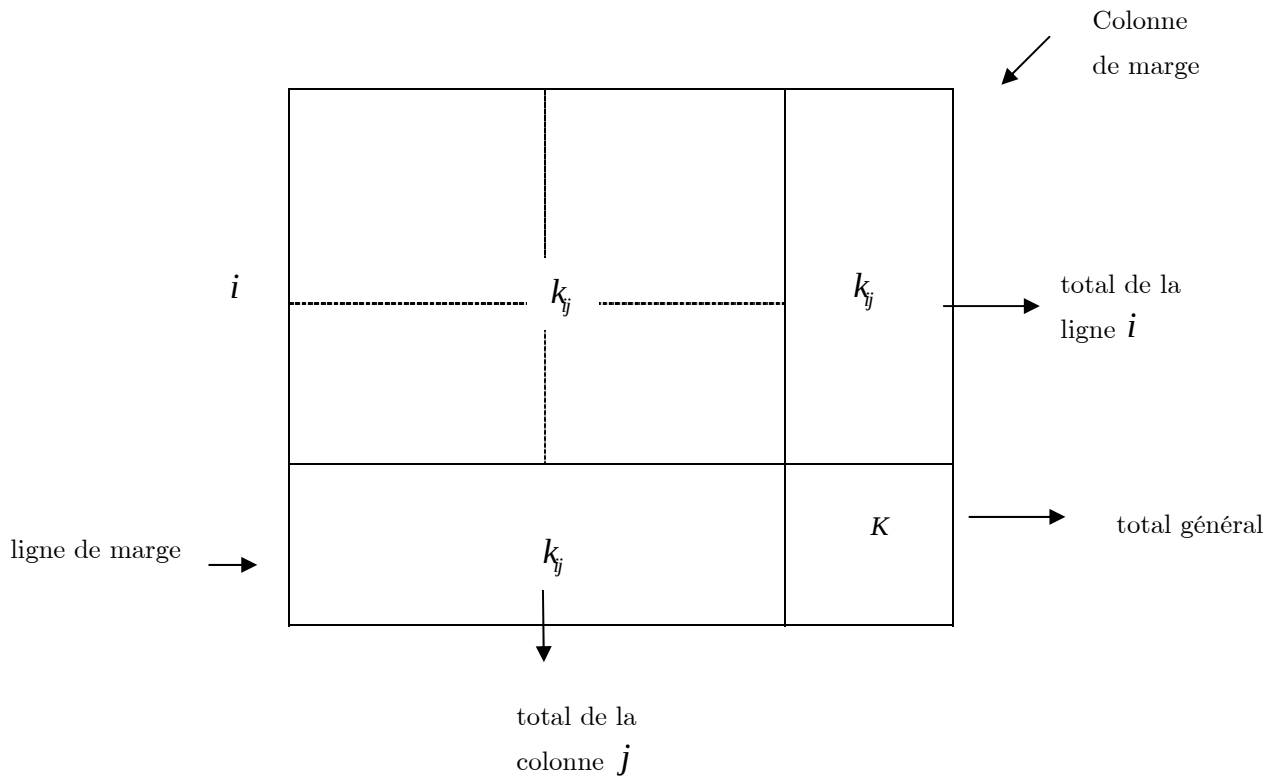
$k_{ij} = \{k_{ij} | i \in I, j \in J\}$, un tableau homogène de nombres sur le produit de ces deux ensembles I et J ($CardI = n, CardJ = p$). On pose :

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

$k_i = \{k_{ij} | j \in J\}$ l'effectif de la ligne i ; la colonne des éléments k_i est la *colonne marginale* ;

$k_j = \{k_{ij} | i \in I\}$ est l'effectif de la colonne j ; la ligne des éléments k_j est la *ligne marginale* ;

$k = \{k_i | i \in I\} = \sum \{k_j | j \in J\} = \sum \{k_{ij} | i \in I, j \in J\}$ est la somme du tableau. On a le schéma suivant :



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

$f_{ij} = k_{ij} / k$: fréquence d'un couple (i, j) .

$f_i = \sum \{f_{ij} | j \in J\} = k_i / k$ est la fréquence d'une ligne i ; la colonne des f_i est la *colonne des fréquences marginales*.

$f_j = \sum \{f_{ij} | i \in I\} = k_j / k$ est la fréquence de la colonne j ; la ligne des f_j est la *ligne des fréquences marginales*.

Par construction on a évidemment :

$$\sum \{f_i | i \in I\} = \sum \{f_j | j \in J\} = 1.$$

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

$f_j^i = \{f_j^i | j=1,...,CardJ\}$, avec $f_j^i = f_{ij} / f_i = k_{ij} / k_i$ et $f_i \neq 0$; $k_i \neq 0$; f_j^i est la *fréquence conditionnelle* du couple (i, j) connaissant i .

$f_i^j = \{f_i^j | i=1,...,CardI\}$ avec $f_i^j = f_{ij} / f_j = k_{ij} / k_j$ et $f_j \neq 0$; $k_j \neq 0$; f_i^j est la *fréquence conditionnelle* du couple (i, j) connaissant j

Le tableau f_j^i correspond au tableau des pourcentages en lignes. C'est donc le profil de la ligne i ; le tableau f_i^j correspond au tableau des pourcentages en colonnes ; on parle alors de profil de la colonne j . On a :

$$\forall i \in I : \sum f_j^i = 1 ; \forall j \in J : \sum f_i^j = 1$$

6.2.2. Analyse des points i de $N_j(I)$ dans R^p

Dans l'espace des colonnes, le point i sera muni de la masse f_i et représenté par son profil f_j^i (sa composante sur la j -ème variable est $f_j^i = f_{ij} / f_i$) ; on notera par :

$$N_j(I) = \{(f_j^i, f_i) | i \in I\}, \text{ le nuage des points } i \in I$$

i) Le centre de gravité, de ce nuage est $f_j = \{f_j; j \in J\}$

Comme le centre de gravité (ou barycentre) G d'un système, de points $\{(m_i, x_{ij}) | i \in I\}$ est le point moyen du système, de la j -ème composante X_{G_j} tel que pour tout j :

$$\sum \{m_i(x_{ij} - x_{G_j}) | i \in I\} = 0$$

On a donc pour tout i :

$$f_i(f_j^i - f_j) - f_i\left(\frac{f_{ij}}{f_i} - f_j\right) = f_{ij} - f_i f_j$$

d'où

$$\sum \{f_i(f_j^i - f_j) | i \in I\} = \sum \{f_{ij} | i \in I\} - f_j \sum \{f_i | i \in I\} = 0$$

étant donné que

$$\sum \{f_{ij} | i \in I\} = f_j, \text{ et } \sum \{f_i | i \in I\} = 1$$

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

$$d^2(i, i') = \left\{ \frac{1}{f_j} (f_j^i - f_j^i)^2 | j \in J \right\}$$

Cette distance mesure les proximités de forme entre lignes (ou entre colonnes) compte tenu de leurs poids différents. Elle est appelée distance du χ^2 (chi-2) et vérifie ce qu'on appelle le principe d'équivalence distributionnelle :

Si deux lignes (ou deux colonnes) du tableau k_{IJ} sont proportionnelles et qu'on les remplace par une seule ligne (ou par une seule colonne) qui en soit la somme colonne par 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 j_1 et j_2 de J tels que leurs profils sur I soient identiques ($f_{i1} = f_{i2}$) ; si on substitue aux colonnes j_1 et j_2 une colonne j_s telle que :

$f_{ij_s} = f_{ij_1} + f_{ij_2}$, $f_{j_s} = f_{j_1} + f_{j_2}$, 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.

iii)- La distance d'un point i au centre de gravité du nuage $N_j(I)$ est :

$$\rho^2(i) = \sum \left\{ \frac{1}{f_j} (f_j^i - f_j)^2 \mid j \in J \right\}$$

iv) De même on peut calculer l'inertie de ce point i caractérisé par son profil f_j^i et par son poids f_i .

On a :

$$I_n(i) = f_i \rho^2(i)$$

v) L'inertie du nuage $N_j(I)$ sera égale à

$$\begin{aligned} I_n(N_j(I)) &= \sum \{ I_n(i) \mid i \in I \} \\ &= \sum \left\{ f_i \sum \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - f_j \right)^2 \mid i \in I \right\} \\ &= \sum \sum \left\{ \frac{f_i}{f_j} \left(\frac{f_{ij}}{f_i} - \frac{f_i f_j}{f_i} \right)^2 \mid i \in I, j \in J \right\} \end{aligned}$$

d'où

$$I_n(N_j(I)) = \sum \sum \left\{ \frac{1}{f_i f_j} (f_{ij} - f_i f_j)^2 \mid i \in I, j \in J \right\}$$

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

On a donc :

$$I_n(N_j(I)) = I_n(N_i(J))$$

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

$\forall j \in J, \forall i \in I$, on associe à f_j^i la quantité $f_j^i f_j^{-1/2}$ alors la distance euclidienne usuelle entre deux points i et i' vaut :

$$d^2(i, i') = \sum (f_j^i f_j^{-1/2} - f_j^{i'} f_j^{-1/2})^2 = \sum \frac{1}{f_j} (f_j^i - f_j^{i'})^2,$$

et on voit qu'elle coïncide bien avec la distance du chi-2. Avec cette transformation le centre de gravité du nouveau nuage que l'on note $N'_j(I) = \{(f_j^i f_j^{-1/2}, f_i) \mid i \in I\}$ est :

$$f_j' = \{f_j^{1/2}; j \in J\}$$

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

$$t_{jj'} = \sum f_i (f_j^i f_j^{-1/2} - f_j^{1/2})(f_{j'}^i f_{j'}^{-1/2} - f_{j'}^{1/2}) ;$$

ce qui conduit à la recherche des vecteurs propres et valeurs propres de la matrice des variances-covariances T qui joue le rôle de $X'X$ dans l'analyse générale.

6.2.3. Analyse des points de $N_i(J)$ dans R^n

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

6.2.4. Relations entre les points i de $N_j(I)$ et les points j de $N_i(J)$

Comme en ACP, les facteurs sont de moyenne nulle et les axes factoriels sont deux à deux orthogonaux (au sens de la métrique du chi-2). On rappelle que l'inertie du nuage $N_j(I)$ projeté sur l'axe α est égale à celle du nuage $N_i(J)$ 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 :

$$F_\alpha(i) = (1/\lambda_\alpha^{1/2}) \sum \{f_j^i G_\alpha(j) | j \in J\} : \text{projection de la ligne } i \text{ sur l'axe de rang } \alpha \text{ de } N_j(I) ;$$

$$G_\alpha(j) = (1/\lambda_\alpha^{1/2}) \sum \{f_j^i F_\alpha(i) | i \in I\} : \text{projection de la colonne } j \text{ sur l'axe de rang } \alpha \text{ de } N_i(J) ;$$

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 f_{ij} est élevée. Sur les plans factoriels, les points éloignés de l'origine, retiennent particulièrement l'attention, car ce sont les profils les plus différents du profil moyen.

Enfin, on peut recalculer les valeurs du tableau initial en fonction des marges et des facteurs. En effet, connaissant les lois marginales f_I et f_J , la suite des facteurs F_α et G_α jusqu'à l'ordre p , et les valeurs propres $\lambda_1, \dots, \lambda_p$, on trouve que :

$$f_{ij} = f_i f_j (1 + \sum \{(1/\lambda_\alpha^{1/2}) F_\alpha(i) G_\alpha(j) | \alpha \in [1, p]\}) ;$$

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

6.2.5- Eléments supplémentaires

Soit i_s une ligne supplémentaire. Pour visualiser i_s sur le α -ème axe factoriel on projette le profil de i_s sur cet axe. L'abscisse $F_\alpha(i_s)$ de cette proposition s'écrit :

$$F_\alpha(i_s) = (1/\lambda_\alpha^{1/2}) \sum \{f_j^{i_s} G_\alpha(j) | j \in J\}$$

De même pour une colonne supplémentaire j_s , l'abscisse $G_\alpha(j_s)$ de la projection du profil j_s sur l'axe α s'écrit : $G_\alpha(j_s) = (1/\lambda_\alpha^{1/2}) \sum \{f_i^{j_s} F_\alpha(i) | i \in I\}$

6.3- Interprétation d'une analyse factorielle des correspondances.

6.3.1. Tableau des données de base

Reprenons notre exemple de le tableau1. Le chef de la scolarité du CASP décide alors de mettre en place un système pour repérer les étudiants en fonction du profil de leurs notes dans les différentes matières concernées. Le fichier à analyser est donc un tableau où chaque étudiant représente une ligne et chaque matière une colonne.

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 $p \leq [\inf(cardI, cardJ) - 1]$; on a : $1 \geq \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. Le vecteur propre associé à la valeur propre 1, est dénommé «vecteur propre trivial» car il n'apporte rien pour l'analyse factorielle de $N_J(I)$ et $N_I(J)$. La première valeur propre à considérer dans notre exemple est en l'occurrence $\lambda_1 = 0.00750$. On a ensuite $\lambda_2 = 0.00515$ et $\lambda_3 = 0.00327$; 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 propres par des longueurs qui leurs sont proportionnelles, ce qui permet d'apprécier d'un regard la décroissance des valeurs propres quand leur rang augmente. Sur notre exemple, on voit que chacune des deux premières valeurs propres est nettement séparée de celle qui la suit : λ_1 vaut près de 1.5 fois λ_2 ; $\lambda_2 = 1.6\lambda_3$. 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 enfin que des valeurs propres élevées "indiquent des oppositions tranchées dont l'interprétation est souvent à la fois évidente et attendue. Des valeurs propres faibles peuvent correspondre à des corrélations plus discrètes que l'analyse aura révélées".

LES VALEURS PROPRES VAL (1) = 1.00000

!NUM ! VAL PROPRE ! POURC.! CUMUL! VARIAT. !*! HISTOGRAMME 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 !	.685 !* ! ***
! 10 !	.00069	! 2.503 !	94.461 !	.643 !* ! ***
! 11 !	.00048	! 1.752 !	96.212 !	.751 !* ! **
! 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.

$$f_i = k_i / k \quad ; \quad f_j = k_j / k$$

ii) L'inertie (INR) qui donne en millièmes la valeur de l'inertie de chaque point f_j^i de $N_j(I)$ (profil de la ligne afférente à l'étudiant i) ou f_i^j de $N_i(J)$ (profil de la colonne afférente à la note j) par rapport au centre de gravité du nuage, rapporté à l'inertie totale du nuage.

$$INR(i) = f_i d^2(i, G) = f_i \rho^2(i)$$

!	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 corrélé avec l'axe et donc bien représenté sur cet axe.

On a ensuite, pour chaque facteur un groupe de trois colonnes.

iv) Le facteur lui-même ; 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 cinq extraits.

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

$$\cos^2 = F_i^2(i) / \rho^2(i)$$

vi) La contribution relative d'un point à un axe (CTR) 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 nuage projeté sur l'axe.

$$CTR = f_i F^2(i) / \lambda_\alpha$$

!	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 et MATH (pour les variables) ; cela ne peut être dû qu'à leur distance élevée à l'origine ;
- les points centraux, pour lesquels *POID* est nettement supérieur à *INR* comme ABDO, BANZ (pour les individus), STAG et APPG (pour les variables) ;
- 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 LOUZ, MBIK et NZAK (pour les individus), DEMO et COME (pour les variables) ;
- les axes qui expliquent l'écart des points au centre de gravité : ce sont ceux pour lesquels *COR* a une forte valeur ;
- Les points bien représentés comme ONDZ, LIK1 et BOYE pour les individus et MATH, COME et DEMO pour les variables.

6.3.4. Représentations graphiques

En AFC, on utilise la représentation simultanée de $N_j(I)$ et $N_j(J)$ 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.

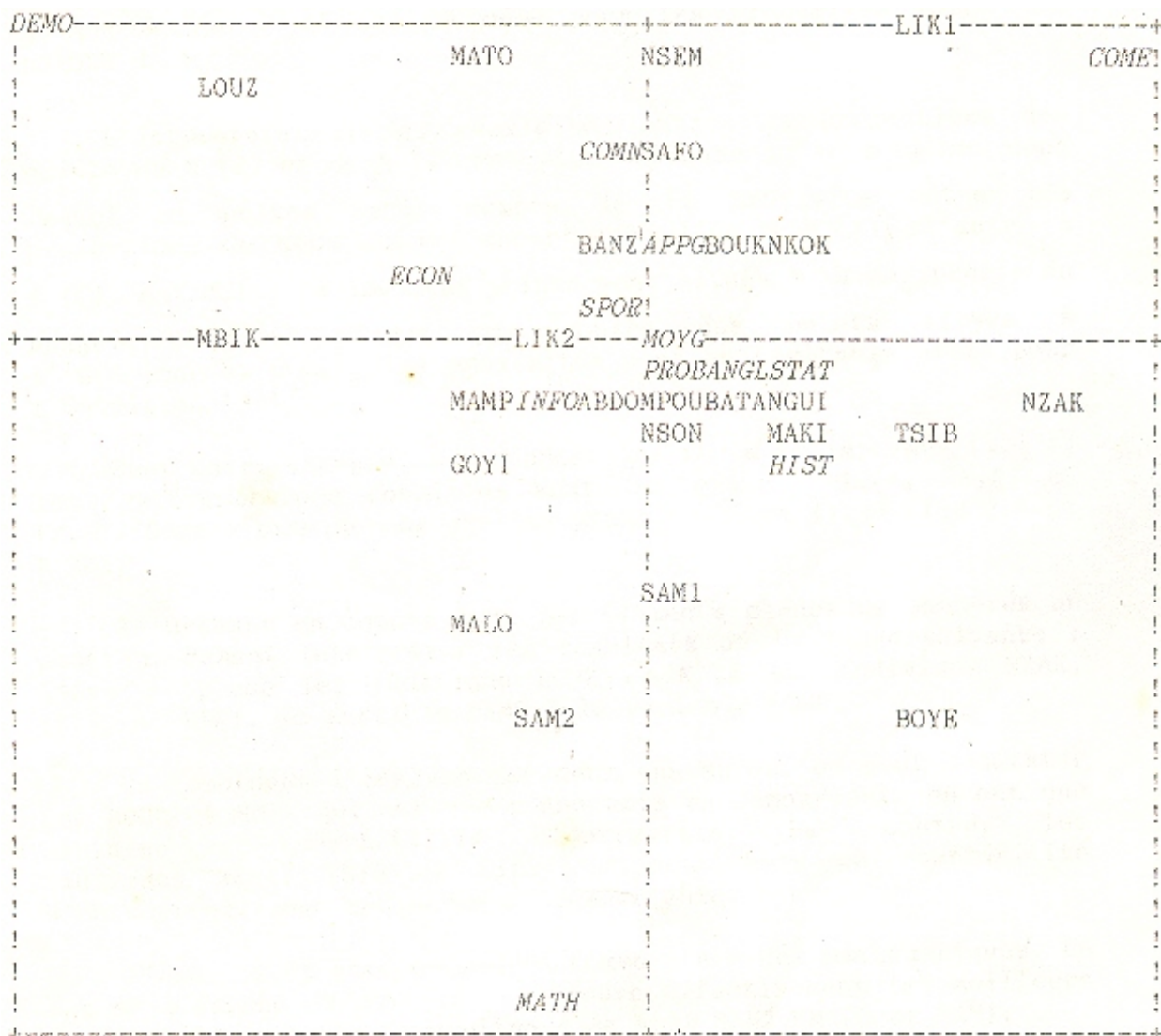
Deux points de $N_j(I)$ proches révèlent un comportement semblable des caractères lignes correspondant pour ces deux axes de projection (il est de même pour les proximités entre les points de $N_j(J)$).

L'interprétation des proximités entre les projections des points de $N_j(I)$ et de $N_j(J)$ est plus délicate ; "le seul cas dans lequel on puisse tenir compte de la proximité entre les projections de deux points appartenant l'un à $N_j(I)$, l'autre à $N_j(J)$, est celui où ces deux points sont situés à la périphérie du nuage... Lorsqu'il s'agit par contre des points situés 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.

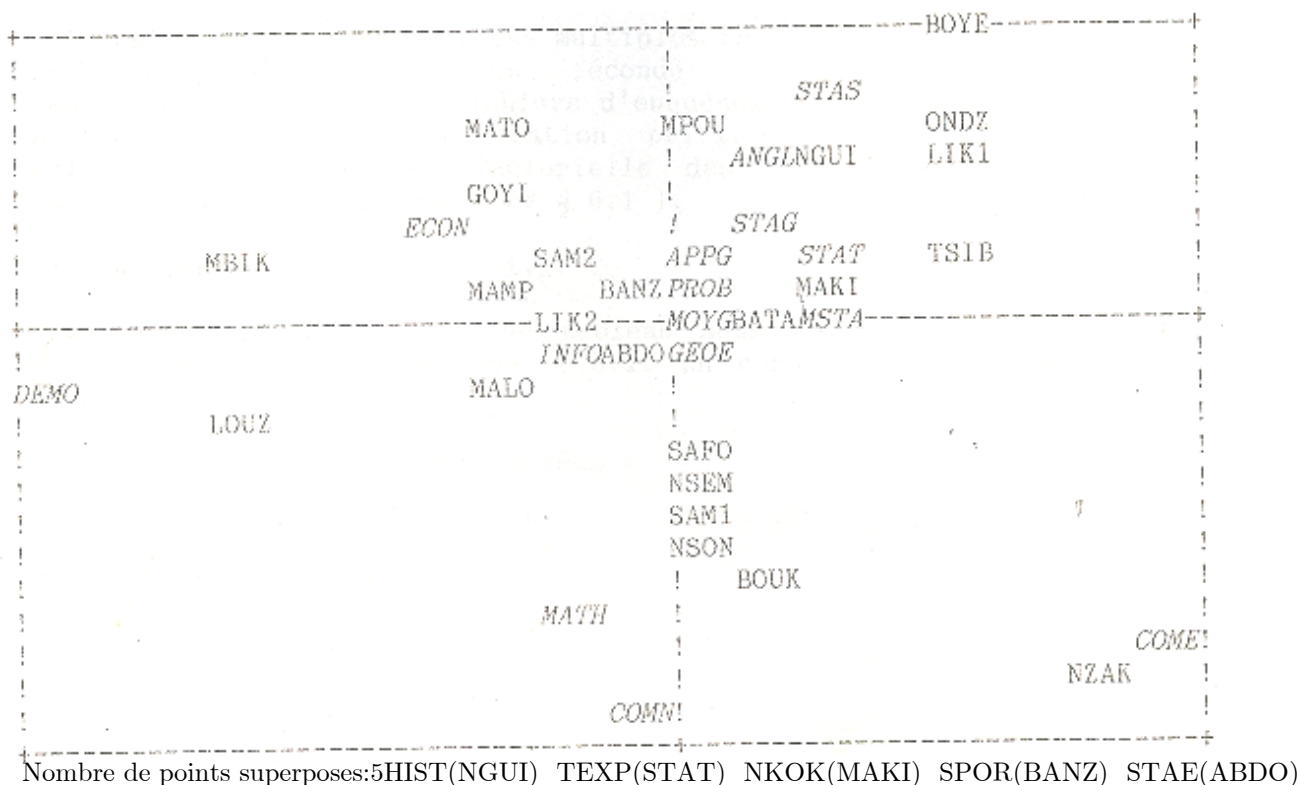
En combinant l'analyse des deux ensembles, on peut constater que LOUZ et MBIK qui ont une bonne note en démographie, et ont une mauvaise note en comptabilité d'entreprise. Par contre, les individus NZAK, TSIB et LIK1 qui sont bons en comptabilité d'entreprise, sont médiocres en démographie.

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



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

Note : Le lecteur peut 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.

6.4 L'analyse des correspondances multiples

6.4.1- Tableau disjonctif complet

Les marges du tableau sont notées :

avec $\sum \{k_{ij} \mid j \in j_q\} = 1$

done

$$\sum \{k_j | j \in J_q\} = Card I = n$$

$$k = \sum \{k_{ij} \mid j \in I, j \in J\} = \sum \{k_i \mid i \in I\} = nCardQ$$

49

On note les fréquences marginales comme suit :

$$f_i = k_i / k = CardQ / nCardQ = 1 / CardI$$

Soit p_j la proportion des individus ayant fourni la réponse j à la question q :

$$P_j = k_j / CardI . (\sum \{p_j | j \in J_q\} = 1) .$$

On a alors : $f_j = p_j / CardQ$

Le nombre d'individus ayant fourni la réponse j est $I \times p_j$; cela veut dire que le chiffre 1 se trouve $I \times p_j$ fois dans la colonne j ; et le chiffre 0 se trouve $I \times (1 - p_j)$ fois.

Le tableau précédent peut être interprété de la manière suivante : dans une enquête comprenant quatre questions, l'individu i a choisi la modalité 1 de la première question (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 modalités de réponses). On vérifie bien que la ligne de marge k_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.

$$\begin{aligned} \forall j \in J \quad , \quad \forall j' \in J : B_{jj'} &= \sum \{k_{ij}k_{ij'} | i \in I\} \\ &= \text{nombre d'individus ayant adopté à la fois les modalités} \\ &\quad j \text{ et } j' . \end{aligned}$$

$$\sum \{B_{jj'} | j' \in J_q\} = k_j$$

$$B_j = \sum \{B_{jj'} | j' \in J\} = k_j CardQ$$

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

	J_1	J_2	J_3	J_4	
J_1	0				
J_2		0			
J_3			0		$B_j = k_j \text{Card} Q$
J_4				0	
					$n(\text{Card} Q)^2$

Il faut faire remarquer que : si j et j' appartiennent au même sous-ensemble j_q de modalités, on a :

$$\forall j, j' \in J_q : B_{jj'} = 0 \text{ si } j \neq j'$$

$$= k_j \text{ si } j = j' \text{ (nombre d'individus ayant adopté la modalité } j \text{)}.$$

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_{IJ} et B_{JJ} définis par :

$$k_{IJ} = \{k_{ij} \mid i \in I, j \in J\} \text{ et}$$

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

avec

$$B_{jj'} = \sum \{k_{ij} k_{ij'} / k_i \mid i \in I\}$$

On a alors :

$$B_j = \sum_j B_{jj'} = \sum_j \sum_i k_{ij} k_{ij'} / k_i = (\sum_i k_{ij} / k_i) \sum_j k_{ij'} = \sum_i k_{ij} = k_j$$

$$B = \sum_{j,j'} B_{jj'} = \sum_j B_j = \sum_j k_j = k$$

Pour procéder à l'analyse factorielle du tableau de Burt, il faut diagonaliser les matrices UU' et $U'U$ étant la matrice de terme général $u_{jj'}$ défini par :

$$u_{jj'} = (B_{jj'} - B_j B_{j'} / B) / (B_j B_{j'})^{1/2}$$

En tenant compte du fait que

$$B_j = k_j ; B_{j'} = k_{j'} ; B = k :$$

$$u_{jj'} = \left[\sum \{k_{ij}k_{ij'} / k_i\} - k_j k_{j'} / k \right] / (k_j k_{j'})^{1/2}$$

On a donc $u_{jj'} = t_{jj'}$ = terme général de la matrice de variances-covariances T du §6.2.2.

Ainsi l'analyse factorielle de B_{JJ} revient à diagonaliser la matrice T^2 .

L'analyse des correspondances de k_{JJ} fournit les mêmes facteurs que celle du tableau B_{JJ} mais, les valeurs propres correspondantes sont différentes : à la valeur propre λ de l'analyse de k_{JJ} correspond la valeur propre λ^2 de l'analyse de B_{JJ} .

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 R^n :

$$\rho^2(j) = d^2(j, G) = \sum \{1/f_i ((f_{ij}/f_j)^2 - f_i) | i \in I\}$$

Comme : $f_i = 1/\text{Card}I$ et $f_j = p_j / \text{Card}I$

on a en définitive :

$$\rho^2(j) = \sum \{ \text{Card}I \left(\frac{\text{Card}Q}{p_j} f_{ij} - \frac{1}{\text{Card}I} \right)^2 | i \in I \}$$

On peut décomposer cette somme selon les valeurs prises par f_{ij} ; on trouve que :

$$f_{ij} = \frac{1}{\text{Card}Q \times \text{Card}I} (\text{Card}I \times p_j) \text{ fois et } 0 \text{ Card}I \times (1 - p_j) \text{ fois ; ce qui donne :}$$

$$\rho^2(j) = p_j \left(\frac{1 - p_j}{p_j} \right)^2 + (1 - p_j)$$

$$\rho^2(j) = (1 - p_j) p_j$$

ii) La contribution de la modalité j vaut donc :

$$CTR(j) = f_j d^2(j, G) = (p_j / \text{Card}Q) d^2(j, G)$$

$$CTR(j) = (1 - p_j) / \text{Card}Q$$

L'inertie due à une modalité est d'autant plus grande 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 contribution d'une question q est :

$$CTR(q) = \sum \{ (1 - p_j) / \text{Card}Q | j \in j_q \}$$

Comme $\sum \{ p_j | j \in j_q \} = 1$ on a :

$$CTR(q) = (\text{Card}j_q / \text{Card}Q) - 1$$

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 :

$$I_n(N_I(J)) = (\text{Card}J / \text{Card}Q) - 1$$

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

6.4.5- Interprétation d'une analyse des correspondances multiples.

6.4.5.1- Tableau des données de base.

A partir du tableau des variables quantitatives (tableau1), on peut construire un tableau 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 bornes des classes.

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 $19 \times 3 = 57$ 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 k_{IJ} 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 précède, six variables seront positionnées en éléments supplémentaires. Ce sont : PROB, ANGL, HIST, SPOR, APPG et MOYG. Le tableau 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 ACM chaque valeur propre est inférieure ou égale à 1 et leur somme est égale à l'inertie totale du nuage, soit : $(CardJ - CardQ) / CardQ = (39 - 13) / 13 = 2$ dans notre exemple. Dans ces conditions, aucune valeur propre ne peut représenter plus que 100/inertie totale, soit : $100 \cdot CardQ / (CardJ - CardQ)$

LES VALEURS PROPRES VAL (1) = 1.00000						HISTOGRAMME DES VALEURS PROPRES !	
NUM !	VAL PROPRE !	POURC !	CUMUL !	VARIAT. !	!		
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 !	* !		

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 $N(I)$ et j de $N(J)$ sont données par les 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 facteurs sur J :

- 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 $CardJ_q = 3$, les contributions à l'inertie de toutes les questions sont égales : $CTR(q) = (CardJ_q - 1) / CardQ = 2/13 = 0.154$

!	I1	!	Q1T	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 !			

Tableau 13 : Facteurs sur I

L'inertie d'une question $INR(q)$ la question q ramenée à l'inertie totale vaut donc :

$INR(q) = CTR(j_q) / \text{inertie totale} = 0,154/2=0,077$ (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 !					

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.

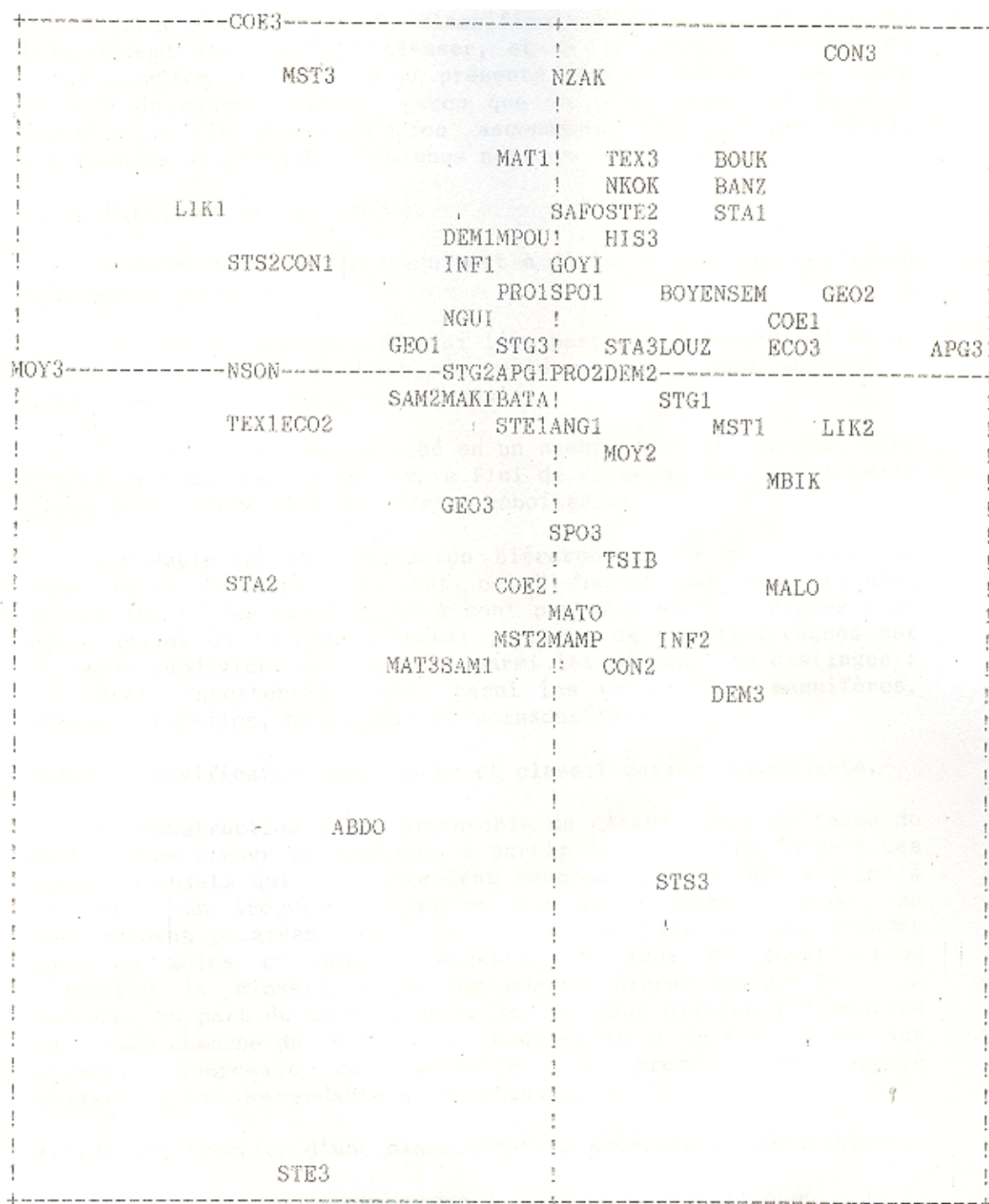
L'axe 3 est dominé par les valeurs élevées des modalités de certaines variables. Ce sont, du côté positif de l'axe MST3, CON3 et COE3 et du côté négatif STS3, STE3 et MAT3. L'opposition de ces variables sur cet axe constitue le trait dominant. On peut, comme pour l'axe 1, tirer les conclusions similaires 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* emboîtées. L'exemple de classification hiérarchique le 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 poissons"[5].

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.

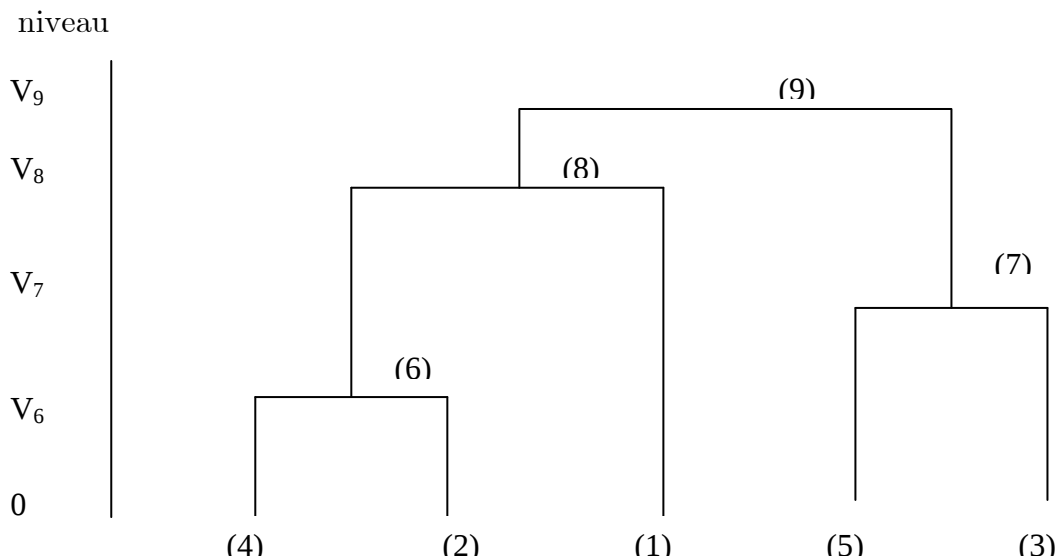
7.1.3- Construction d'une classification ascendante hiérarchique.

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 v_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)).



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 $2CardI - 1$.

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}) :

entre les ensembles de points q et q' $d_{\text{saut}}(q, q')$ est la distance minima entre un point de q et un point de q' . 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}) :

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

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

$d_{\text{moy}}(q, q')$ est la moyenne des distances entre un point de q et un point de q' . 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 l'ensemble I est considéré comme un nuage de 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 ; $N_j(I)$ le nuage des éléments i de I et affectés de masse m_i . On rappelle qu'une inertie est le produit d'une masse par le carré d'une distance.

- L'inertie du nuage $N_j(I)$ s'écrit :

$$\begin{aligned} I_n(N_j(I)) &= \sum \{I_n(i) | i \in I\} \\ &= \sum \{m_i \rho^2(i) | i \in I\} \end{aligned}$$

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

Soit q une partie de I , on notera par m_q sa masse totale, et G_q ou simplement q son centre de gravité.

- L'inertie d'une classe q s'écrit :

$$I_n(q) = \sum \{m_i d^2(i, q) | i \in q\} ;$$

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

$$\begin{aligned} I_n(Q) &= \sum \{I_n(q) | q \in Q\} \\ &= \sum \{m_q d^2(q, G) | q \in Q\} ; \quad m_q = \sum \{m_i | i \in q\} \end{aligned}$$

A toute partition de I en un ensemble Q de classes q correspondant une décomposition de l'inertie du nuage $N_j(I)$ en inertie interclasses et inerties intra-classes suivant la formule (relation de Huygens) :

$$I_n(N_J(I)) = I_n(Q) + \sum \{I_n(q) | q \in Q\}$$

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 m_a et m_b que l'on agrège en une seule classe n de Q_n de masse $m_a + m_b$. Le critère qui conduit aux choix de a et b est celui qui rend minimum la perte d'inertie réalisée en passant de Q_{n-1} à Q_n :

$$\nu(n) = [I_n(Q_{n-1}) - I_n(Q_n)] \text{ minimum.}$$

Ce qui équivaut à maximiser l'inertie de la partition Q_n (ou encore à maximiser le moment centré d'ordre 2 de la partition). A chaque pas, on minimise l'inertie intra-classe de la partition construite. La quantité $\nu(n)$ est également appelée indice de niveau (cf. § 7.1.3).

$$\nu(n) = \frac{m_a \cdot m_b}{m_a + m_b} d^2(a, b)$$

L'inertie totale de la classe n peut selon la relation de Huygens être décomposée en inertie des deux classes a et b dont la réunion est n et $\nu(n)$, terme proportionnel au carré de la distance entre aîné et benjamin du nœud n :

$$I_n = I_a + I_b + \nu(n)$$

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

$$\sum \{\nu(n) | n \in N\} = I_n(N_J(I)).$$

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 I des étudiants. On rappelle quelques formules du §6.2.1 :

- la masse de l'élément i de I : $m_i = f_i = k_i / k$;

- la distance de chi-2 entre profils :

$$d^2(i, i') = \sum \{1 / f_j (f_j^i - f_j^{i'})^2 | j \in J\}$$

- l'inertie de l'élément i :

$$I_n(i) = f_i \sum \{1 / f_j (f_j^i - f_j)^2 | j \in J\}$$

- la masse de la classe q :

$$f_q = \sum \{f_i | i \in q\}$$

- le profil de la classe q :

$$f_j^q = \frac{f_{q_j}}{f_q} = \frac{\sum f_{ij} | i \in q}{f_q}$$

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

$$I_n(q) = f_q \rho^2(q) ; \text{ avec } \rho^2(q) = \sum \{1/ f_j (f_j^q - f_j)^2 | j \in J\}$$

- l'indice de niveau du nœud n :

$$\nu(n) = \frac{f_a f_b}{f_a + f_b} d^2(a, b) = \sum \{1/ f_j (f_j^a - f_j^b)^2 | j \in J\}$$

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 le taux d'inertie cumulé $T(Q)$.

SOMME DES INDICES DE NIVEAU .29124 E-01

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

Avec le critère de l'inertie adoptée, la somme des indices de niveau est égale à l'inertie totale du nuage des individus.

! J !	I(J) !	A(J) !	B(J) !	P(J) !	DESCRIPTION DES CLASSES DE LA HIERARCHIE									
! 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 est très forte, ceci symbolise le fait qu'il n'existe que quelques séparations principales. Les niveaux les plus bas de la hiérarchie peuvent être considérés comme des intermédiaires de 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

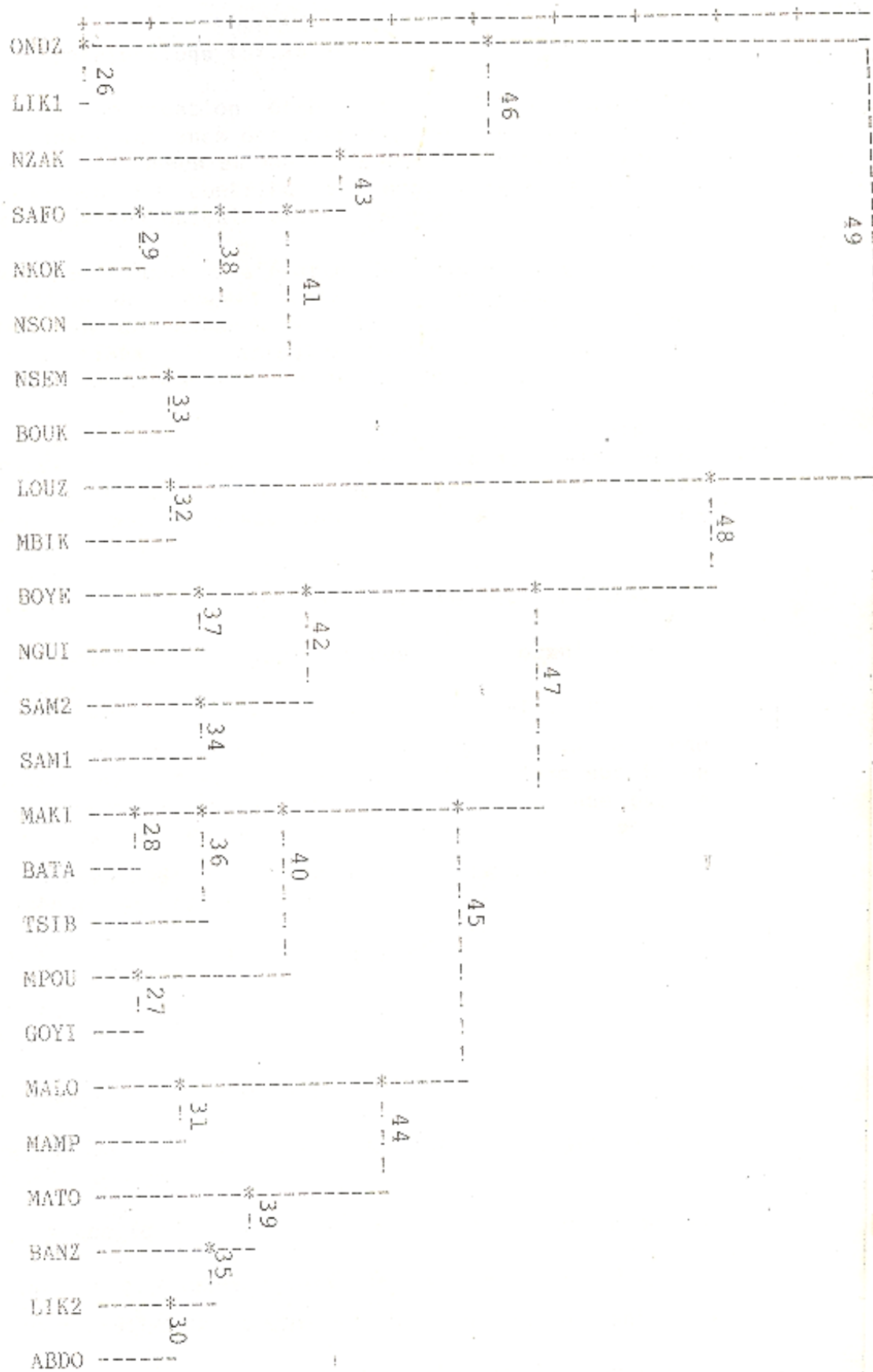
On a construit $CardI - 1$ classes, c'est-à-dire $25 - 1 = 24$ classes. Les classes de la hiérarchie sont numérotées de 26 à 49. Chaque classe est décrite par : son numéro J , son indice de niveau $I(J)$, ses successeurs $A(J)$ et $B(J)$, le nombre de ses éléments $P(J)$ et la liste des éléments de chaque 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 $A(40) = 36$ (MAKI, BATA et TSIB), puis les deux éléments de $B(40) = 27$ (MPOU et 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 q s'écarte du centre de gravité du nuage : c'est l'étude de 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 $b(n)$ en lesquels se scinde la classe n . Comme précédemment, cette étude sera réalisée d'abord dans l'espace rapporté au système d'axes factoriels, ensuite dans l'espace rapporté au système des axes définis par les variables ;
- les contributions mutuelles entre classes et facteurs.

7.2.5.1 Etude des classes par rapport à des axes. Formulaire

Soit k_{IJ} , 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 f_q 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 : $F_\alpha(q) = \sum \{f_i F_\alpha(i) / f_q \mid i \in q\}$, avec

$$f_q = \sum \{f_i \mid i \in q\}$$

Pour une classe q , on calcule :

- le cosinus carré de l'angle formé entre le rayon vecteur de la classe q et l'axe α :

$COR_\alpha(q) = \cos^2_\alpha(q) = F_\alpha^2(q) / \rho^2(q)$. C'est la contribution relative de l'axe α à l'excentricité de la classe q , ($\rho(q) = \rho^2(q)$) étant l'excentricité de la classe q par rapport au centre du nuage).

$COR_\alpha(q) \neq 0 \Rightarrow \alpha$ intervient peu dans l'écart de q au centre du nuage.

$COR_\alpha(q) \neq 1 \Rightarrow \alpha$ explique l'écart de q au centre du nuage.

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

$$\sum COR = QLT$$

- la contribution relative de la classe q au facteur α : $CTR_\alpha(q) = f_q F_\alpha^2(\alpha) / \lambda_\alpha$.

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

$CTR_\alpha(q) \neq 0 \Rightarrow q$ n'explique pas l'inertie de l'axe α ;

$CTR_{\alpha}(q) \neq 1 \Rightarrow \alpha$ 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. On a extrait cinq facteurs ; seuls les deux premiers facteurs sont indiqués. Pour le premier tableau la première ligne n'est pas utilisable (sauf la colonne poids) puisque le $\rho^2(q)$ est égal à zéro. Le poids total du nuage étant normalisé à 1, le centre du nuage 49 à pour poids 1000/1000=1.

AXES FACTORIELS 1 A 2

(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000)

CLASSE	AIN	BNJMN !	POIDS	INR	QLT !	F 1	COR	CTR !	F 2	COR	CTR
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). Pour le reste des lignes, l'interprétation se base sur la définition des formules ci-dessus indiquées comme en ACP ou en AFC.

AXES FACTORIELS 1 A 2

(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000)

CLASSE	AIN	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 ;
- par contre, l'inertie relativement à l'origine 0 du centre d'une classe q 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 pourquoi le total de la colonne *INR* est inférieur à 1 ; ce total représente l'inertie interclasse de la partition retenue (ici exprimé en millième) et l'inertie intra-classes, le complémentaire à 1 de ce total. De façon analogue la somme des *CTR* 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 q au centre de gravité du nuage. On se place donc ici dans l'espace des profils sur l'ensemble des *CardJ* variables : à chaque variable j correspond un axe, la 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 STAS, MATH, GEOE, etc. (profil sur J de la ligne de marge du tableau k_{JJ}). 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 dira 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 lisant la colonne *COR*. 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, $QLT = 1000$. Le fait de n'avoir retenu que quelques variables, cette valeur de la qualité de représentation est descendue au dessous de 1000 et ce, pour toutes les classes. On vérifie que les valeurs des colonnes *POIDS* et *INR* sont les mêmes que dans le tableau 17. On peut faire éditer les résultats similaires pour les centres de gravité des six classes de la partition, définies à partir des cinq nœuds les plus hauts.

(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000, A L'EXCEPTION DE RHO2 QUI EST MULTIPLIE PAR 10** (5))

CLASSE	AINE	BNJM !	POIDS	INR	QLT	RHO2!	STAS	COR	CTR!	MATH	COR	CTR
49	46	48 !	1000	0	0	0!	64	0	0!	58	0	0
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	RHO2 !	TEXP	COR	CTR !
49	46	48 !	1000	0	0	0 !	51	0	0 !
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.

A chaque nœud n d'une classification, est associé un dipôle $(a(n), b(n))$ 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 $a(n), b(n)$ dans les dipôles. On calcule :

- la différence $D_{\alpha}(n) = F_{\alpha}(a(n)) - F_{\alpha}(b(n))$; elle renseigne sur la position relative de $a(n)$ par rapport à $b(n)$.

- $COD_{\alpha}(n) = (F_{\alpha}(a(n)) - F_{\alpha}(b(n)))^2 / \|a(n) - b(n)\|^2$
 $= D_{\alpha}^2(n) / \|a(n) - b(n)\|^2$

$$= Q_n D_n^2(n) / \nu(n)$$

avec : $Q_n = f_{a(n)} f_{b(n)} / f_n$

$COD_\alpha(n)$ est le cosinus carré de l'angle formé par l'axe α et la droite joignant les centres des classes $a(n)$ et $b(n)$:

si $COD_\alpha(n) \neq 1 \Rightarrow \alpha$ explique en quasi totalité la séparation entre $a(n)$ et $b(n)$;

si $COD_\alpha(n) \neq 0 \Rightarrow \alpha$ n'explique pas la séparation entre $a(n)$ et $b(n)$

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

$$CTD = Q_n D_\alpha^2(n) / \lambda_\alpha$$

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

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

L'inertie totale du nuage sur l'axe 1 est expliqué à 29% ($CTD(48)=291$) par la dichotomie entre les deux classes $A(48)=32$ et $B(48)=47$ formant la classe 48. En projection sur l'axe 1, la partition de I en deux classes : 32 et 47 a une inertie interclasse de : $291 \times \lambda_1$ (et une inertie intra-classe de $709 \times \lambda_1$). On vérifie que le total de la colonne CTD du tableau 20 est égal à celui de la colonne CTR du tableau 18 (classes de la partition). Il en est de même pour le total de la colonne IND (total IND =total INR =504).

AXES FACTORIELS 1 A 2

(TOUTES LES VALEURS SONT MULTIPLIES PAR 1000)

NŒUD	AINE	BNJMN !	POIDS	IND	QLD !	D 1	COD	CTD !	D 2	COD	CTD
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 *D2AB* : c'est le carré de la distance entre les centres de classe *a(n)* et *b(n)*.

Pour l'interprétation, on cherche à repérer les valeurs élevées de la colonne *COD*. Le dipôle (46,48) est expliqué par la variable *COME* (comptabilité d'entreprise) ; le dipôle (32,47) est expliqué par STAS (Statistique de la santé) ; le dipôle (26,43) par MATH (mathématiques) ; le reste des dipôles n'a pas d'explications particulières.

TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000, A L'EXCEPTION DE D2AB QUI EST MULTIPLIE PAR 10**(4)

NŒUD	AINE	BNJMN !	POIDS	IND	QLD	D2AB!	STAS	COD	CTD!	MATH	COD	CTD
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	COD	CTD
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 $I_n(N_J(I))$ ou par $M^2 N_J(I)$ l'inertie totale du nuage. On sait déjà que :

$$I_n(N_J(I)) = M^2(N_J(I)) = \sum \{v(n) | n \in N\} \\ = \sum \{\lambda_\alpha | \alpha \in A\}$$

(N ensemble des nœuds et A ensemble des facteurs)

$$M^2(N_J(I)) = \sum \{v(n; \alpha) | \alpha \in A; n \in N\}$$

avec

$$v(n; \alpha) = Q_n D_\alpha^2(n)$$

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

On peut aussi noter que :

$$\lambda_\alpha = \sum \{v(n; \alpha) | n \in N\}$$

$$v(n) = \sum \{v(n; \alpha) | \alpha \in A\}$$

On peut enfin calculer les contributions relatives mutuelles entre classes et facteurs. C'est le rapport :

$$v(n; \alpha) / M^2(N_J(I))$$

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.

Dans le tableau 21, on a limité le nombre des variables dans l'étude des dipôles. Cette étude est ici complétée (tableau 23) par le tableau des contributions mutuelles relatives entre dipôles et variables. La dernière ligne donne pour chaque variable l'inertie relative de $M_J^2(I)$ 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.

Dans l'espace factoriel on peut représenter, soit les fourches (le triplet n , $a(n)$, $b(n)$ 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

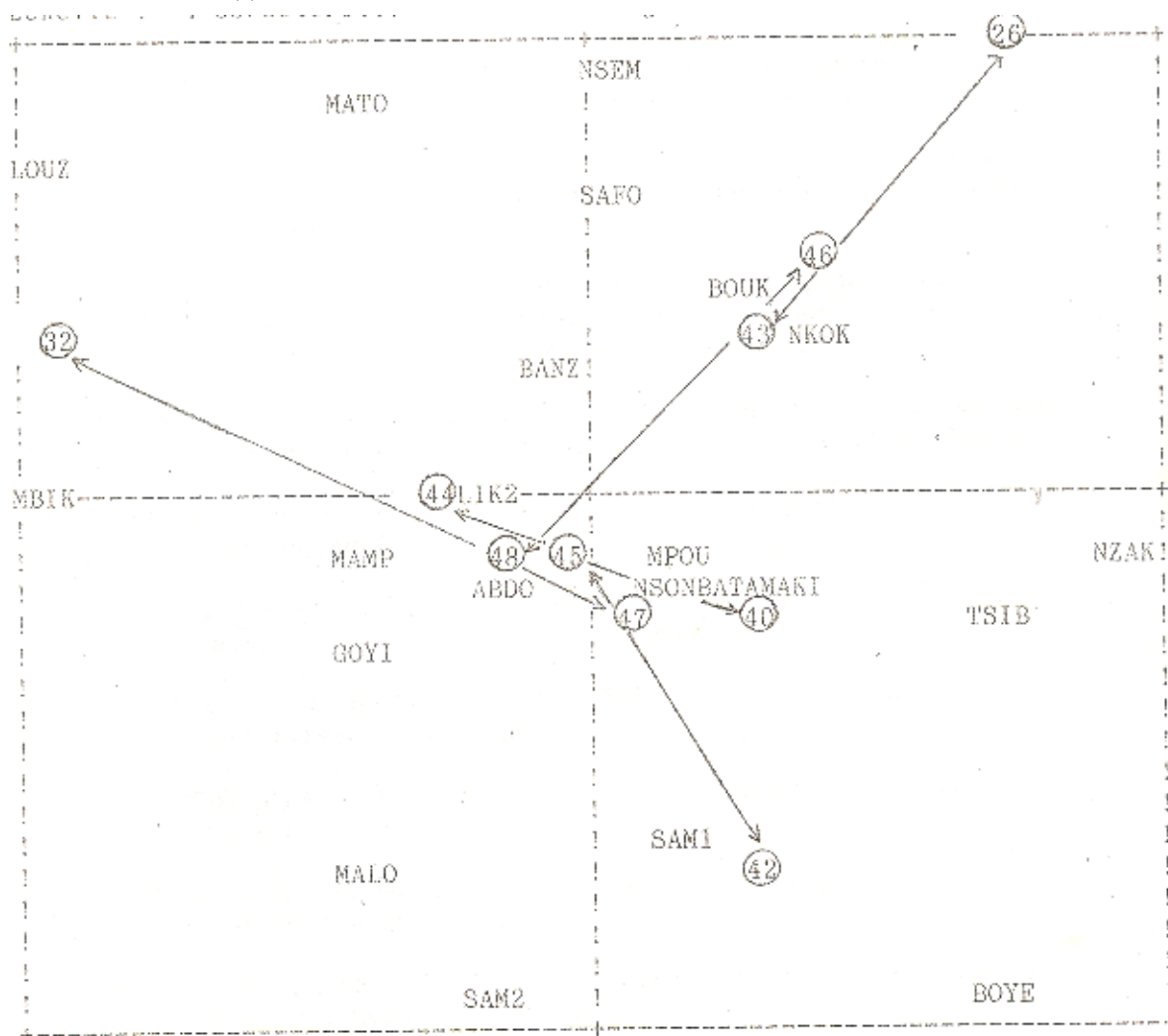
		1	2	3	4	5
N	26	119	124	67	-75	-16
N	27	-35	-30	67	78	-41
N	28	44	-27	8	62	-15
N	29	26	60	-13	7	-50
N	30	-43	-13	-7	-25	64
N	31	-85	-59	-3	-47	-47
N	32	-196	44	-7	12	-34
N	33	16	82	-69	63	35
N	34	-16	-115	-42	-49	47
N	35	-37	0	-2	5	67
N	36	69	-31	13	42	11
N	37	77	-83	79	-22	16
N	38	18	31	-38	-17	-51
N	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). *Analyse 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.