CENTRE DE RECHERCHES
ET DE DOCUMENTATION
SUR LA CONSOMMATION

COMITE D'ORGANISATION ET
DE RECHERCHE APPLIQUEE
SUR LE DEVELOPPEMENT
ECONOMIQUE ET SOCIAL

RECHERCHES SUR LA DESCRIPTION AUTOMATIQUE

DES DONNEES SOCIO-ECONOMIQUES

1973

Centre de Recherches et de Documentation
sur la Consommation

Comité d'Organisation et Recherche
Appliquée sur le Développement
Economique et Social

Convention de Recherche n° 13/1971

RECHERCHES SUR LA DESCRIPTION AUTOMATIQUE DES DONNEES SOCIO-ECONOMIQUES

L. LEBART et N. TABARD / cd

21 MARS 1973 - N° 4172

S O M M A I R E

RESUME

Ce rapport comprend quatre chapitres principaux, un chapitre d'introduction, une annexe.

Le chapitre 0 (d'introduction) tente de situer en termes généraux les techniques d'analyse de données dans l'ensemble des techniques statistiques, et insiste sur les aspects spécifiques de ces méthodes et leur contribution à certaines recherches socio-économiques. Cette introduction fait souvent référence au chapitre 4, qui est un recueil de trois exemples pratiques d'application.

L'analyse des données apparaît comme un nouvel *instrument d'observation* répondant à la nécessité *d'organiser sans intervention* certains matériaux statistiques complexes, sans qu'il soit nécessaire de réduire ou de simplifier a priori le champ de l'observable.

On insiste surtout sur les qualités nécessaires du recueil de données, sur la mise en évidence de faits statistiques, sur la technique de régression visualisée (projection de variables illustratives), particulièrement utile lors des dépouillement d'enquêtes.

Le chapitre 1, *"Exposé technique des méthodes"*, reprend un cours professé en partie à l'I.S.U.P. (Cycle de Statistique appliquée et Centre d'enseignement et de recherche en statistique appliquée) et dans le cadre d'un groupe de travail sur l'analyse des données à l'E.N.S.A.E. Cet exposé nous a semblé nécessaire pour donner une certaine autonomie à l'ensemble de ce rapport, puisque les notions qu'il introduit sont supposées connues dans les chapitres suivants. Il s'agit d'un exposé technique mais élémentaire. Des considérations plus historiques que pratiques nous font commencer par l'analyse en composantes principales. L'analyse des correspondances est présentée de divers points de vue. Suivent ensuite, pour information et à titre d'exercice, l'analyse factorielle classique, les analyses canoniques et discriminantes.

Les chapitres 2 et 3 sont deux contributions à l'étude des données socio-économiques.

Le chapitre 2 traite de *l'analyse de certaines correspondances multiples*. On propose, pour l'étude de ce type de données, une méthode qui s'introduit naturellement à partir de l'étude des questionnaires mis sous forme disjonctive complète (formés de questions dont les modalités s'excluent mutuellement).

Les programmes d'application en FØRTRAN IV commentés, sont adaptés aux fichiers volumineux. Ils comportent notamment un programme de construction de partition adapté à ce type de questionnaire. Un exemple d'application figure en fin de chapitre : "Description de la structure de l'échantillon d'une enquête".

(L'exemple 1 du chapitre 4 constitue également une application de la méthode évoquée dans ce chapitre).

Le chapitre 3, sous le titre : *Description de certaines relations binaires*, comporte deux volets : dans une première partie, on étudie la façon dont l'analyse des correspondances restitue des structures élémentaires telles que celle de graphe symétrique. Le seconde partie concerne les relations binaires connues de façon exogène entre individus statistiques, telles que les relations de voisinages existant entre des zones géographiques. On définit une technique d'"analyse locale" permettant de mettre éventuellement en évidence l'échelle des relations entre variables. On propose également une généralisation de l'analyse discriminante. Le listage des programmes en FØRTRAN IV est suivi par un exemple d'application à l'étude de la structure socio-professionnelle des 80 quartiers de Paris.

Le chapitre 4, rédigé par N. TABARD, comporte trois exemples d'application.

Le premier exemple traite de l'analyse des questionnaires. On éprouve par l'analyse des données la cohérence des réponses à des questions d'attitude sur le thème de la politique des prestations familiales (enquête CNAF 1971). L'analyse fait ressortir les principales significations de ces réponses : opposition entre prestations en espèces (allocations diverses) et prestations en nature (services et équipements collectifs) d'une part et d'autre part réactions à l'institution éventuelle de critères sélectifs pour l'attribution des prestations. L'analyse des données permet dans ce cas une interprétation des non-réponses.

Le second exemple porte sur des budgets familiaux. On traite comme un tableau de contingence l'ensemble des dépenses de consommation des ménages de la communauté économique européenne. Les ménages sont répartis en 55 groupes socio-géographiques, les dépenses en 65 postes de consommation. La carte géographique de la C.E.E. ressort dans le plan des deux premiers facteurs.

Le troisième exemple est l'analyse d'un tableau de contingence reposant sur un recensement partiel : la répartition des familles allocataires résidant dans les communes-échantillon de l'enquête CNAF 1971, selon le nombre d'enfants et la perception ou non de l'allocation de salaire unique (ces trois critères croisés).

Enfin, on a jugé utile de donner en annexe, rédigée avec la collaboration de G. LACOURLY, des listages des programmes d'analyse de données usuels en langage A P L.

CHAPITRE 0

CHAPITRE 0

REMARQUES SUR L'APPORT DES METHODES D'ANALYSE DES DONNEES A LA RECHERCHE EN SCIENCES HUMAINES

Dans ce chapitre d'introduction nous allons résumer le point de vue d'un praticien, non spécialiste des questions d'épistémologie des sciences de l'homme, sur l'apport éventuel des techniques d'analyse des données à la recherche socio-économique. Nous allons, pour cela, être amené à définir ces techniques en insistant dans la mesure du possible sur ce qui peut être considéré comme original et acquis. Il va de soi qu'un même outil peut donner lieu à toutes sortes d'usages et de pratiques, et que l'originalité ou le progrès à attendre des méthodes que nous étudions seront également tributaires de la démarche intellectuelle du chercheur.

Nous devrons également dire quelques mots de la recherche socio-économique, en essayant de nous limiter à la toute petite portion de cette recherche qui nous intéresse plus directement : la connaissance des faits.

Nous nous appuierons d'une part sur des exemples concrets d'applications de ces méthodes, d'autre part sur des écrits, dont la responsabilité de l'interprétation incombe à l'auteur de ces remarques.

I - L'analyse des données : définition pratique et généralités.

Le traitement statistique d'un tableau de valeurs numériques pourra être considéré comme relevant de l'analyse des données, si toutes les conditions ci-dessous sont satisfaites simultanément :

a) le tableau à étudier est vaste, ou au moins suffisamment important pour que sa taille soit un obstacle à une connaissance directe de son contenu.
b) il n'existe aucun modèle a priori, de nature probabiliste par exemple, concernant les éléments du tableau (ce qui ne signifie pas que la méthode d'analyse ne puisse éventuellement s'inspirer de la nature même du tableau).
c) le traitement se fait à l'aide d'un ordinateur.

Ces trois caractéristiques ne sont pas indépendantes : plus le tableau à étudier comprendra de variables, plus le champ des hypothèses probabilistes réalistes se rétrécira ; de plus, le volume des calculs impliquera le recours à

un traitement automatique. Ainsi, la première des caractéristiques impliquera souvent les deux suivantes.

Une dénomination équivalente de l'analyse des données est "Statistique descriptive multidimensionnelle". Par contre, l'expression "Analyse multidimensionnelle" désigne plus généralement tous les traitements statistiques mettant en jeu de nombreuses variables, y compris les techniques impliquant l'usage d'hypothèses statistiques extrêmement fortes, telle que l'analyse multidimensionnelle de la variance ; (il en est de même pour les expressions "analyse multivariée" ou "analyse multivariate").

Nous ferons donc cette distinction terminologique importante, qui semble admise par les principaux fondateurs de l'analyse des données.

Ainsi défini, l'ensemble des techniques d'analyse des données comprend principalement :
- les techniques d'analyse factorielle descriptive.
- les techniques de classification automatique.

I-1. L'analyse factorielle descriptive.

Le principe général de cette méthode, ainsi que ses diversifications, notamment l'analyse des correspondances, est exposé en détail (bien que de façon élémentaire) au chapitre I.

En résumé, tout tableau de nombres de dimensions (n x p) peut donner lieu, après d'éventuelles transformations, à la construction de deux nuages de points dans des espaces ayant respectivement n et p dimensions. Les techniques d'analyse factorielle descriptive ont pour but de réduire les dimensions de ces espaces (en ajustant aux nuages des sous-espaces vectoriels de faibles dimensions) afin d'obtenir des représentations visuelles des proximités existant entre les points des nuages.

L'analyse des correspondances joue un rôle privilégié pour l'étude des tableaux homogènes de nombres positifs, car les hypothèses mises en jeu sont assez faibles, l'interprétation des proximités est facile, enfin, les propriétés d'invariance sont satisfaisantes. (Symétrie entre les deux dimensions du tableau, caractérisation des points des nuages par des profils, propriétés de la distance).

Pour utiliser une image que nous reprendrons dans le chapitre consacré à la description des relations binaires, l'analyse factorielle descriptive est un outil qui fonctionne un peu comme un appareil radiographique ou un microscope électronique.

On obtient des représentations à partir d'un phénomène caché ou inaccessible, et l'on induit la structure ou certains traits pertinents du phénomène à partir de règles d'interprétation. L'idéal serait de disposer d'une technique "standard" afin de pouvoir en banaliser l'usage, et de permettre au plus grand nombre de chercheurs de l'utiliser. C'est là un peu le rôle dévolu à l'analyse des correspondances dans la plupart des applications en socio-économie, notamment en raison des propriétés favorables de cette méthode rappelées ci-dessus.

I-2. La classification automatique

Ces méthodes ont pour objet de fournir des représentations assimilables à partir de tableaux rectangulaires de données, mais en faisant apparaître des structures telles que des partitions, des hiérarchies de partitions, des arbres (structures mettant en jeu des relations d'inclusion, d'ordre, d'incidence).

On peut classer ces méthodes de classification d'après les grands principes techniques mis en oeuvre, ou au contraire du point de vue de l'utilisateur.

Schématiquement, il existe deux grandes catégories de méthodes :
- les méthodes dites "ascendantes" (ou agglomératives), qui partent des individus à classer et les agrègent progressivement.
- les méthodes dites "descendantes", qui segmentent l'ensemble des individus, en construisant de proche en proche des classes de plus en plus fines.

Non moins schématiquement, il existe deux grands types d'utilisations de ces méthodes (qui ne sont pas forcément disjoints au cours d'une même application). Il s'agit de :
- découvrir des structures préexistantes
- organiser les données.

Pour désigner cette dernière démarche, M.G. KENDALL parle quelquefois de "dissection" qu'il oppose ainsi à classification.

Ces deux démarches possibles concernent d'ailleurs également l'analyse factorielle pour laquelle cette dichotomie est cependant moins tranchée, de

par le caractère continu des représentations obtenues.

En fait, nous parlerons assez peu des méthodes de classification automatique, qui, du point de vue des applications au domaine socio-économique, n'apportent en général pas d'avantages substantiels en sus des méthodes d'analyse factorielle. Toutefois nous serons amenés à utiliser des méthodes de recherche globale de partitions sur certains fichiers de grandes dimensions (cf. chapitre 3), en utilisant des algorithmes qui ne sont pas les plus élégants du point de vue théorique (méthode "quick and dirty", selon les statisticiens anglo-saxons), mais qui sont les seuls utilisables sur les fichiers volumineux qui nous intéressent principalement. Les méthodes de classification hiérarchique peuvent rendre de grands services dans les problèmes de constructions de nomenclatures. Cependant, la multiplicité de ces techniques et l'effervescence qui règne autour de ce domaine de recherche tout nouveau nous ont conduits à ne retenir que les techniques dont l'utilité est incontestée : comme le signale un statisticien[1] il est presque plus facile actuellement pour un chercheur d'inventer à propos d'un recueil de données un nouvel algorithme de classification, que de tirer quelque chose de ces données !.... Cette boutade a le mérite de rappeler que tout travail interdisciplinaire est menacé par la fuite dans la technique de certains de ses participants qui cèdent alors à cette force centrifuge pour éviter l'ascèse d'une communication toujours difficile.

II - Exhibition et illustration d'une structure

Il est fréquent, lors de l'analyse d'un tableau de données, d'illustrer les typologies obtenues en projetant des variables supplémentaires. Contrairement aux variables ayant servi à exhiber ces typologies à partir du corpus, ces variables illustratives peuvent concerner les thèmes les plus variés, et constituer un ensemble assez hétérogène... Ainsi, la typologie des communes de la région parisienne[2] a pu être illustrée successivement par une mise en évidence de l'âge des immeubles, du nombre d'enfants, des équipements sanitaires des logements, par projection des différentes catégories socio-professionnelles avec individualisation des sexes, etc... En somme la typologie obtenue est une trame qui peut recevoir différents tissages suivant le thème retenu.

Cette procédure d'illustration n'est autre qu'une régression, où l'on ne s'intéresse pas principalement aux "coefficients de régression", comme cela est courant lors de la plupart des utilisations de ces techniques, mais à

---

1 - Cf. R.M. CORMACK - A review of classification. Applied statistics 1971 - Rapport de synthèse et discussion.

2 - L. LEBART et N. TABARD (op. cité).

certains aspects descriptifs que nous allons expliciter brièvement.

Plaçons nous, comme cela est fait classiquement, dans l'espace dont les coordonnées représentent des observations. Les points de cet espace sont des variables, que nous supposerons dichotomisées en variables exogènes (ou explicatives) et endogènes (ou variables à expliquer). Sommairement, disons que les variables exogènes engendrent une certaine variété linéaire, sur laquelle on projettera les différentes variables endogènes (que l'on désire expliquer séparément par ces variables exogènes). Les coordonnées de cette projection dans la base formée par les variables exogènes sont précisément les coefficients de régression. On sait que si cette base est une "mauvaise base", c'est-à-dire s'il existe des colinéarités entre les variables exogènes, ces coefficients seront assez imprécis. Le problème qui nous intéresse ici est assez différent nous aimerions voir, au sein de la variété linéaire que nous qualifierons en bref d'explicative, quelles sont les positions des projections des variables à expliquer vis-à-vis des différentes variables exogènes? Pour que cette vision soit possible, il nous faut d'abord chercher un ajustement par un sous-espace à deux dimensions de cette variété, puis projeter les points représentant les variables endogènes sur ce sous-espace : on reconnait là une variante de la régression linéaire multiple (régression après orthonormalisation des variables exogènes), qui coïncide, dans cette optique d'utilisation, avec l'illustration de la typologie des variables exogènes par les variables endogènes.

On peut toujours dans cette optique, considérer une analyse de données comme la première étape d'une régression qui prépare notre corpus à accueillir les variables illustratives.

Cette technique de régression visualisée jour un grand rôle lors des dépouillements d'enquêtes. Ainsi, dans l'exemple du chapitre II, § VI, la structure de l'échantillon d'après les caractéristiques objectives des ménages constitue un "résumé explicatif potentiel" sur lequel les variables les plus diverses pourront être projetées.

Aux fastidieux dépouillements des tableaux croisant des variables deux à deux est substituée une représentation qui utilise l'ensemble des dépendances entre variables. Si le "résumé explicatif potentiel" est constitué à partir de l'ensemble des modalités de 30 variables, la projection des modalités d'une variable supplémentaire nous fournit une représentation remplaçant la lecture de 30 nouveaux tableaux croisés.

Le gain n'est pas seulement mesurable arithmétiquement, car la représentation obtenue est chargée de significations qui la rendent vivante et assimilable.

Cette méthode de régression visualisée ne permet pas seulement de procéder à des dépouillements rapides, qui sont autant de coloriages de la carte structurale qu'il y a de variables illustratives ; elle permet de vérifier la cohérence logique d'un système de réponses et de détecter d'éventuelles erreurs ou anomalies.

Ce seul aspect de l'utilisation des méthodes suffirait à justifier leur emploi systématique dans les dépouillements d'enquêtes.

Considérons ainsi l'exemple 1 du chapitre IV. Un questionnaire d'enquête laissait la possibilité aux enquêtés de ne pas répondre à une question, alors que les modalités prévues couvraient en fait toutes les réponses possibles.

Les non-réponses traduisent donc un comportement de l'enquêté que l'on peut qualifier en bref d'anormal, ou de déviant. Les effectifs de non-réponses sont d'ailleurs faibles. Comme le phénomène de "refus de répondre" est distinct du thème qui fait l'objet des questions, l'analyse n'a été faite que sur les modalités correspondant à une réponse effective (les non-réponses ayant été ventilées de façon aléatoire pour chaque individu entre les diverses modalités de réponses de la question correspondante). La projection des non-réponses, dans un second temps, permet de voir, pour chaque question, de quelle modalité de réponse effective se rapprochent les non-réponses, et donc de comprendre le phénomène des non-réponses qui apparaissent dans ce cas particulier notamment comme des substitut des réponses faisant intervenir des négations.

III - Qualités nécessaires d'un recueil de données

Nous avons utilisé jusqu'ici l'expression "tableau de données" comme s'il s'agissait de quelque chose de parfaitement défini. En fait, n'importe quel tableau ne sera pas analysable, si certaines conditions générales ne sont pas réalisées. L'analyse d'un tableau de données fournit des résultats ayant une certaine permanence, et, l'expérience le montre, pouvant être interprétés aisément, si ce tableau de données vérifie les deux conditions d'homogénéité et d'exhaustivité (1).

---

(1) cf. J.P. BENZECRI : (1970) La pratique de l'analyse des correspondances, in "L'analyse des données" Tome IIA n°2 - DUNOD 1973.

Un tableau de contingence, croisant deux partitions d'une même population, possède ces deux qualités de façon idéale : les mesures sont des fréquences absolues d'individus ou d'occurrences diverses , elles sont toutes de même nature et il est licite de calculer des sommes sur les lignes et les colonnes du tableau.

Le critère d'homogénéité est donc satisfait.

Le critère d'exhaustivité est lui aussi satisfait puisque nous avons affaire à des partitions (tout individu appartient à une classe, et nous prenons en compte toutes les classes). Il s'agit bien entendu ici d'exhaustivité vis-à-vis d'une population particulière. En fait, le concept d'exhaustivité se rapporte à un phénomène, et la population étudiée, si elle ne résulte pas d'un recensement global, devra de plus être dans une certaine mesure représentative du phénomène. Le mot "représentative" est pris ici dans un sens beaucoup plus large que dans la théorie des sondages, et beaucoup plus imprécis (en l'absence de travaux théoriques permettant d'affiner cette notion) : la population doit recouvrir tous les aspects du phénomène, sans qu'une importance excessive soit attribuée à un système de pondération ou de stratification. (Seuls, les moments du second ordre interviennent de façon directe dans les calculs).

La condition d'homogénéité permettra de considérer les lignes (ou les colonnes) du tableau de données comme les éléments d'un même espace, et de définir entre ces éléments une distance qui soit la moins conventionnelle possible. La condition d'exhaustivité permet de délimiter par des frontières qui soit les plus naturelles possibles les nuages construits à partir de ces éléments.

Si la définition de l'homogénéité est relativement claire (l'addition des éléments des lignes ou des colonnes du tableau doit avoir un sens), la notion d'exhaustivité reste vague, lors de nombreuses applications, car elle suppose une définition rigoureuse du "phénomène" observé, ce qui est parfois difficile a priori. Lors de l'étude de la structure socio-professionnelle des 80 quartiers de Paris (cf. chapitre III, § 2.6), le tableau de contingence analysé semble avoir toutes les qualités requises ; cependant, le fait de se limiter à l'étude de la population "intra-muros", en coupant Paris de l'ensemble de l'agglomération donne des frontières assez conventionnelles à l'étude. A posteriori, pourtant, il apparaîtra que le boulevard périphérique n'est pas une limite si conventionnelle, et que la population de Paris mérite d'être étudiée de façon

isolée. Une étude de l'ensemble des communes de la région parisienne[1] nous a prouvé que, à quelques exceptions près, les quartiers de Paris formaient un sous-ensemble assez facilement isolable. En fait, les deux qualités que nous exigeons des tableaux de données font partie des recommandations générales que les linguistes s'imposent pour constuire un corpus ; ainsi, à propos de la recherche sémiologique[2] Roland BARTHES écrit : "D'une part, le corpus doit être assez large pour qu'on puisse raisonnablement espérer que ses éléments saturent un système complet de ressemblances et de différences ; il est sûr que lorsque l'on dépouille une suite de matériaux, au bout d'un certain temps on finit par rencontrer des faits et des rapports déjà repérés"..."ces retours sont de plus en plus fréquents jusqu'à ce qu'on ne retrouve plus aucun matériau nouveau : le corpus est alors saturé. D'autre part, le corpus doit être aussi homogène que possible ; d'abord, homogénéité de la substance ; on a évidemment intérêt à travailler sur des matériaux constitués par une seule et même substance, à l'instar du linguiste qui n'a affaire qu'à la substance phonique..."

En fait, la meilleure justification des caractéristiques du recueil de données est celle qu'impose une méthode de travail, assez générale, qui consiste, lors de l'étude de systèmes complexes, à ne pas mêler les hypothèses et les conclusions : une typologie construite à partir de mesures non homogènes ou de relevés partiels est en effet difficile à interpréter : elle est hypothéquée d'une part par les arbitraires dans les codages inhérents à l'hétérogénéité des mesures, d'autre part, par le troncage du champ global des observations. Comment faire la part des manipulations et des carences dans la lecture des résultats ? Lorsque les variables sont très nombreuses, on peut vérifier, par des simulations adaptées, que les arbitraires de codages ne jouent éventuellement qu'un rôle limité... grâce aux grands nombres, l'exhaustivité peut ainsi venir au secours de l'hétérogénéité. En fait, dans ce dernier cas, il existe toujours des possibilités de codages homogènes redonnant les mêmes résultats (analyse des rangs[3], codages binaires disjonctifs, etc.)

On pourra vérifier que pour les cinq exemples figurant dans ce rapport, les recueils de données satisfont bien ces recommandations :

1) - le cas des quartiers de Paris a déjà été rapidement examiné.
2) - l'exemple de la structure de l'échantillon de l'enquête CNAF-CREDOC (chapitre II, § 6) est peut-être le plus délicat à justifier de ce point de vue. La dimension "individus" du tableau est bien homogène et exhaus-

---

1 - Morphologie sociale des communes urbaines. L. LEBART et N. TABARD - Consommation 1971.

2 - R. BARTHES - Elements de sémiologie - (GONTHIER) - 1964.

3 - L. LEBART - J.P. FENELON - Statistique et Informatique appliquées - DUNOD - 2ème édition - 1973.

tive de par la construction de l'échantillon ; la dimension "variable", si elle est homogène du point de vue du codage (questionnaire disjonctif complet) ne l'est pas du point de vue de la signification et de l'importance des variables (qui sont des caractéristiques objectives très diverses des ménages). Disons que le principe de pertinence, recommandé par les linguistes[1], qui consiste à ne retenir dans la masse hétérogène des faits que ceux se rapportant à un seul point de vue, n'est absolument pas suivi dans ce type d'application, dont la finalité est précisément d'obtenir une photographie de l'ensemble des ménages enquêtés dans le foisonnement de leurs situations. L'échantillonnage de ces situations est heureusement suffisamment riche pour que la représentation obtenue puisse prétendre à une certaine stabilité.

3) - l'exemple des attitudes vis-à-vis de la politique familiale (chapitre IV - § 1) réalisé à partir du même échantillon d'individus que l'exemple 2, traite un ensemble de questions dont le codage est homogène, qui se rapportent toutes à un thème précis, et dont il est permis de penser qu'elles "saturent" ce thème.

4) - dans l'exemple des structures de consommation des socio-régions de la C.E.E. (chapitre IV - § 2), la dimension "consommation" est une partition en 67 postes de l'ensemble des dépenses des ménages ; elle vérifie donc les deux principes de façon parfaite. La dimension "socio-région" est un peu plus hétérogène, puisque les résultats relatifs à une même région et à deux catégories de salariés distinctes forment deux unités statistiques distinctes, au même titre que deux régions. Le fait d'avoir limité l'enquête aux régions de la C.E.E. et à deux catégories de salariés peut également apparaître comme une troncature d'un phénomène plus général (la C.E.E., en 1963, ne constitue pas un système clos ou autonome). Toutefois, ces objections de détail, comme l'analyse elle-même le prouve, n'invalident pas la qualité du recueil de données.

5) - l'exemple de la structure des villes selon la composition des familles et le travail des femmes (chapitre IV - § 3) vérifie bien les deux recommandations puisqu'il s'agit d'un tableau de contingence. En fait, l'une des partitions analysées provient du croisement de deux partitions : selon que la mère travaille ou pas, selon la composition de la famille. Le principe de pertinence évoqué plus haut n'est donc pas respecté. Homogène dans sa forme, le tableau ne l'est pas dans son contenu. Cette particularité est bien entendu

---

1 - A. MARTINET - Eléments de linguistique générale - A. COLIN - 1960.

exhibée par l'analyse elle-même, où la première bissectrice des axes factoriels est un axe "nombre d'enfants" et la seconde un axe "travail de la mère" ; ces deux partitions sont assez indépendantes entre elles, mais aucune des deux n'est indépendante de la partition en ville : le premier facteur décrit en réalité l'intéraction des trois types de modalités : activité de la mère, nombre d'enfants, ville .

IV - Accumuler des faits ou des idées

L'histoire des sciences montre que l'observation systématique des faits et l'expérimentation ont toujours accusé un retard sur la déduction. Comme l'explique fort bien PIAGET[1], dont nous résumerons brièvement l'argumentation, ce retard est exceptionnellement important dans le cas des sciences humaines.

L'expérimentation suppose une dépense d'énergie, une organisation, une soumission à des instances extérieures sans rapport avec les conditions d'une réflexion déductive. De plus, le réel est toujours d'une grande complexité (il faut savoir pour voir...), et la lecture d'une expérience suppose déjà acquis un certain cadre logico-mathématique, si fruste soit-il. A ces premières sources de retard, valables pour n'importe quelle science, s'ajoutent, dans le cas des sciences humaines, des obstacles qui tiennent dans l'ensemble au caractère flou de la frontière du sujet égocentrique et sujet épistémique : limitons nous aux domaines sociologiques et économiques qui nous intéressent plus directement.

Les observations systématiques, ou, plus rarement, les expériences sont différées ou jugées peu utiles de par la richesse des intuitions immédiates possibles. Des données ou des faits qui ne sont qu'anecdotiques, l'expérience personnelle du sujet permettent parfois de mettre en route des mécanismes spéculatifs auprès desquels un rassemblement méthodique des faits peut paraître un piétinement stérile.

Ces mécanismes spéculatifs peuvent d'ailleurs utiliser des cadres logico-mathématiques très élaborés, leur cohérence étant alors parfois un substitut de leur adéquation à la réalité.

En sociologie, l'objet de la connaissance est un "nous" collectif dans lequel le sujet est impliqué par différence. Les faits eux-mêmes sont le produit d'un certain découpage réel, lui-même tributaire du système de valeurs du sujet. Ainsi toujours d'après PIAGET (op. cité page 31) "La décentration

---

1 - J. PIAGET - Epistémologie des sciences de l'homme - GALLIMARD - UNESCO - 1970.

comparatiste est en ce cas si difficile que ROUSSEAU, pour penser le phénomène social en cherchant ses références dans les comportements élémentaires et non civilisés (ce qui marquait un grand progrès par rapport aux idées de son temps) imagine le bon sauvage comme un individu antérieur à toute société, mais en lui prêtant, sans s'en rendre compte, tous les caractères de moralité, rationalité, et même de déduction juridique que la sociologie nous apprend être les produits de la vie collective. Ce bon sauvage est même le produit d'une imagination si peu décentrée qu'il ressemble étonnamment à ROUSSEAU lui-même..."

Le rassemblement des faits, pouvant donner lieu à des mesures, à des comptages ou à des codages particuliers constitue, pour n'importe quelle discipline une phase ingrate du travail de recherche. De plus la difficulté d'une décentration du sujet dans les sciences sociales fait peser sur les recueils de données la menace d'une circularité qui peut paraître dangereuse, et semble conduire éventuellement à des tautologies.

Des exemples concrets d'application vont nous permettre de voir quelles peuvent être les contributions des techniques d'analyse des données à la résolution de ce problème central, et de remarquer "qu'un mouvement circulaire qui se ressère pour mieux cerner la réalité n'est pas un cercle vicieux"[1]

L'idée générale est que, un peu comme en micro-physique ou l'expérimentateur modifie le phénomène qu'il veut observer, l'observable se situe à un niveau intermédiaire entre le phénomène et l'observateur ; entre les faits, qui résultent d'un certain découpage du réel hypothèqué par les valeurs ou les normes du sujet épistémique, et les idées ou les intuitions que celui-ci a du phénomène, existent parfois des invariants, qui peuvent se dégager à partir d'une synthèse automatique.

Nous appellerons ces invariants des "faits statistiques".

La notion de fait statistique :

On peut dans un premier temps définir un fait statistique comme l'apparition d'une structure, ou plus simplement d'une description, ayant des propriétés d'invariance et de permanence, à partir d'une synthèse automatique (et donc reproductible) d'un tableau de données statistiques vérifiant les deux propriétés d'homogénéité et d'exhaustivité.

---

1 - J.P. BENZECRI - Analyse des données et mesure des grandeurs - note multigraphiée.

La statistique descriptive multidimensionnelle n'est pas qu'une simplification ou une illustration de la réalité, comme l'était la statistique descriptive classique (histogramme, diagrammes triangulaires, lissage rudimentaire de séries chronologiques), dans la mesure où les résumés qu'elle introduit sont plus "découverts" que construits. L'idée de découverte s'impose progressivement lors des épreuves destinées à éprouver la stabilité des résultats des analyses. Le fait statistique mis en évidence ne constitue cependant qu'un matériau, parfois plus fiable[1] que les matériaux qui ont servi à le construire pris isolément, qu'il convient d'étudier à nouveau, en utilisant des informations extérieures au corpus de données analysées (cf. par exemple le rôle des variables illustratives évoqué plus haut).

Prenons l'exemple de l'étude de la morphologie sociale des communes de la région parisienne, ou celle des quartiers de Paris (synthétisée par la figure 1 page 170), qui donne une typologie des professions et activités du même type.

La typologie obtenue est largement indépendante des nomenclatures utilisées, parce que celles-ci sont assez fines (qu'il s'agisse de la nomenclature en catégories d'activité ou en zones géographiques), de par les propriétés mêmes de la méthode de réduction utilisée.

Le premier axe factoriel extrait (d'une supériorité écrasante du point de vue de la variance expliquée) s'interprète aisément en termes de "statut social du lieu de résidence". Cet axe est conservé si l'on agrège les 80 quartiers en 20 arrondissements, du point de vue de la cohabitation entre les diverses professions, et il est également conservé si l'on agrège les 29 catégories d'activité en une dizaine de postes.

Le concept de statut social, attaché à un lieu de résidence ou à une profession,existe ; il n'est pas question de l'annexer à titre définitif en l'enfermant dans une définition analytique. Il est cependant possible de faire jaillir cette notion à partir de relevés statistiques et de programmes de calcul, qui peuvent les uns et les autres être mis en oeuvre de façon analogue à propos d'autres champs d'application : autre agglomération que la région parisienne, autre nomenclature des catégories d'activité, etc...

Le progrès dans la connaissance réside bien dans la possibilité de répéter et donc de communiquer certains types d'expériences.

---

1 - De par les propriétés de stabilité des résultats vis-à-vis des codages, des nomenclatures, des échantillonnages de variables.

La résurgence de certaines formes constitue en quelque sorte une confirmation de l'existence de faits à un niveau qui n'est pas celui des données brutes : l'exemple de la typologie des régions du marché commun à partir de leurs profils de consommation (chapitre IV - § 2) nous montre qu'un nuage de points-régions, dont les coordonnées sont des postes de dépenses, admet pour meilleure approximation plane la carte géographique de ces régions, sans qu'aucune information d'ordre géographique sur les positions des régions ait été introduite. Ici, le fait statistique coïncide avec une donnée de fait : la disposition des régions les unes par rapport aux autres.

L'aspect graphique des informations issues des analyses de données permet d'ailleurs de façon assez inattendue, des progrès assez nets dans la perception de certains ensembles complexes de faits socio-économiques : le langage usuel, de par son caractère linéaire et séquentiel, privilégie dans la description des relations les liaisons non symétriques du type "implication" (la notion de causalité est plus facile à exprimer que la notion apparemment plus naïve de covariation). Le langage graphique est également plus neutre parce qu'antérieur au choix d'un style qui véhiculera un système de connotation.

L'invariance des résultats peut faire l'objet de vérifications empiriques à propros de chaque application particulière : on peut vérifier, lors d'une enquête par questionnaire, que des "ponctions" réalisées de façon aléatoire dans l'échantillon et dans l'ensemble des questions ne modifient pas de façon substantielle les représentations obtenues ; que celles-ci ne dépendent pas d'éventuelles erreurs de mesure, de codage, de compréhension que l'on est en droit d'attendre de la part des participants et des fabricants de l'enquête.

L'invariance et la stabilité des résultats sont peut-être une occasion de parler de la "double convergence" que l'on observe en analyse des données, et qui distingue ces techniques des méthodes de statistique le plus usuelles. Lorsque nous avons parlé d'exhaustivité, à propros des tableaux de données, nous avons préconisé cette qualité pour les deux dimensions de ces tableaux, qui jouent en général des rôles analogues sinon symétriques.

Le schéma classique de répétition d'épreuves indépendantes ou quasi-indépendantes qui est à la base de la plupart des résultats les plus utilisés de la théorie des sondages ne s'appliquera pas à la lettre (comment parler d'échantillonnage dans un ensemble potentiel de questions qu'il n'est évidemment pas possible de probabiliser ?), mais les principes géométriques de convergence s'appliquent doublement, puisque, de façon idéale, les deux dimensions du tableau sont aussi grandes que possible.

Disons sommairement, que la loi des grands nombres nous dit comment améliorer notre connaissance du niveau d'une variable lorsque le nombre des individus sur lesquels cette variable est mesurée augmente indéfiniment. Mais la connaissance des relations entre variables est améliorée et renforcée lorsque le nombre de variables augmente également.

Si la matrice des covariances se stabilise lorsque la taille de l'échantillon augmente, les premiers axes principaux extraits à partir de cette matrice se stabilisent également lorsque les variables sont nombreuses.

Ceci implique en particulier le fait que les typologies obtenues par ces méthodes à partir d'enquêtes par sondage soient assez indépendantes du redressement de l'échantillon.

Les moments du second ordre sont en effet moins sensibles que ceux du premier ordre aux éventuels systèmes de pondération, et l'espace des premiers facteurs est lui-même peu sensible aux fluctuations éventuelles de ces moments. Ceci est en fait l'ébauche d'une justification du primat de l'exhaustivité (toutes les situations doivent être représentées) sur la représentativité, pour ce type d'analyse.

CHAPITRE I

EXPOSE TECHNIQUE DES METHODES

I - Analyse en composantes principales (ACP)

II - Analyse des correspondances (AC)

III - Analyse factorielle classique

IV - Analyse canonique

V - Analyse discriminante

VI - Contrôle de validité des résultats en analyse factorielle.

I - ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

I-1. GENERALITES

L'analyse en composantes principales est une technique de description de données multidimensionnelles proposée entre les deux guerres par HOTELLING.

Elle est le produit d'une simplification de l'analyse factorielle classique mise au point au début du siècle par des psychologues.

L'apparition des moyens de calcul actuels a permis de diffuser et d'améliorer cette technique, sans toutefois la modifier dans ses principes.

Nous profitons de l'exposé sur l'analyse en composantes principales pour mettre en évidence le noyau théorique commun à toutes les méthodes d'analyse des données usuelles, sous le nom d'Analyse Générale.

I-1.1 - Domaine d'application

La situation pratique devant laquelle se trouve un utilisateur éventuel de cette technique est la suivante : on possède un tableau rectangulaire de mesures, dont les colonnes figurent par exemple des variables (des mensurations, des taux, etc...) et dont les lignes représentent les individus sur lesquels ces variables sont mesurées.

En biométrie, il est fréquent de procéder à de nombreuses mensurations sur certains organes, ou même certains animaux. En économie, il est fréquent que l'on ait à noter les diverses dépenses d'un ménage. Les tableaux numériques ainsi obtenus, dès qu'ils sont un peu volumineux, sont difficiles à synthétiser, et l'information qu'ils contiennent n'est pas assimilable de par son foisonnement même.

La caractéristique que doivent remplir ces tableaux pour être l'objet d'une description par l'ACP est la suivante :

- l'une au moins des dimensions du tableau est formée d'unités ayant

un caractère répétitif, l'autre pouvant être éventuellement plus hétérogène.

Les lignes , dans les exemples cités plus haut, ont ce caractère répétitif : on les désignera en général sous le nom "d'individus" ou "d'observations", les colonnes étant désignées sous le nom de variables Ces lignes peuvent apparaître comme des réalisations indépendantes de vecteurs aléatoires, dont les composantes sont les différentes variables.

NOTATIONS : Le tableau de données sera désigné par la lettre X en notation matricielle. La matrice X sera d'ordre np, autrement dit, elle aura n lignes et p colonnes. Son terme générique est x_{ij}, i-ème observation de la j-ème variable ; dans tous les développements qui vont suivre, n sera l'effectif des individus, p le nombre de variables. La transposée de X sera notée $\tilde{X}$, elle a donc p lignes et n colonnes.

I-1.2 - Interprétations géométriques.

Pour comprendre les opérations de réduction pratiquées par l'analyse en composantes principales, il est utile de représenter géométriquement les lignes et les colonnes du tableau X par des "points" d'un espace à p ou n dimensions.

Les n lignes peuvent être considérées comme n points d'un espace à p dimensions, noté R^p, alors que les p colonnes peuvent également être considérées comme des points d'un espace à n dimensions noté R^n.

- Dans R^p, les proximités existant entre les points (qui représentent des individus ou des observations) ont une interprétation directe pour l'utilisateur : si deux points sont très proches, cela signifie que dans l'ensemble, leurs p coordonnées sont très proches, donc que ces deux individus sont caractérisés par des variables voisines. (Par exemple, les deux ménages représentés par ces points consomment les mêmes produits, de façon analogue).

- Dans R^n, la proximité de deux points-variables signifie que pour ces deux variables, les n individus ont des valeurs voisines. Cela peut vouloir dire que ces variables mesurent une même chose,

ou encore sont très liées.

Toutefois, l'interprétation de ces proximités est évidemment très fruste pour l'instant : des problèmes d'échelles de mesure se posent d'emblée : comment calculer effectivement la distance de deux variables si l'une est exprimée en centimètres et l'autre en kilogrammes ? Comment interpréter une proximité moyenne dans R^p ? Est-ce que deux individus moyennement proches dans R^p ont des valeurs "moyennement proches" pour chacune des variables, ou au contraire très proches pour certaines et éloignées pour d'autres ?

L'ACP nous permettra de répondre à ces questions ; supposons provisoirement que ces proximités ont un sens : le principe général de la méthode va alors consister à les représenter graphiquement par ajustement. On cherchera, dans R^p et R^n, des sous-espaces vectoriels de faibles dimensions (une, deux, trois par exemple) qui ajustent au mieux les nuages de points-individus et de points-variables, de façon à ce que les proximités projetées reflètent, autant que faire se pourra, les proximités réelles.

La mise en oeuvre de l'ACP nécessite en première approximation deux types de considérations, et utilise deux types de résultats : les uns sont purement mathématiques, les autres sont de nature statistique.

Nous allons nous efforcer de séparer ces deux types de démarches et de résultats, afin que l'utilisateur éventuel sache clairement distinguer ce qui est définitivement acquis de ce qui nécessitera de sa part prudence, finesse et circonspection.

Le paragraphe suivant exhibe donc, sous la forme d'un problème d'approximation numérique, les résultats mathématiques sous-jacents à l'ACP.

La lecture du paragraphe III n'est pas nécessaire à la compréhension de l'ACP. Elle le sera par contre pour l'étude de l'Analyse des correspondances et de l'analyse discriminante. Le paragraphe IV reviendra sur les aspects statistiques de l'ACP.

I-2. ANALYSE GENERALE

Nous désignerons comme précédemment par X un tableau (à n lignes et p colonnes) de valeurs numériques. Supposons que l'on sache, a priori, que le tableau X contient des informations redondantes ou dépendantes les unes des autres, et que l'on se propose de réduire le volume de stockage du tableau X ; ainsi, le tableau X peut avoir 2 000 lignes et 300 colonnes, et représenter les réponses à 300 questions de 2 000 individus constituant un échantillon statistique.

La matrice X a ainsi 600 000 éléments. Pour des raisons diverses, il existera des liaisons fonctionnelles ou stochastiques entre certaines questions. Peut-on résumer ces 600 000 valeurs par un nombre inférieur de valeurs sans perte notable d'information, compte tenu de ces liaisons et interrelations?

Il existe, dans certains cas très particuliers, des solutions empiriques à ce problème de réduction du codage de l'information. Si les questions sont codées en 0 ou 1 selon l'absence ou la présence d'un caractère, la réponse 0 (non) à la question "Etes-vous marié ?" impliquera la réponse 0 à la question "Votre femme a-t-elle plus de 40 ans ?", d'où, dans ce cas, une réduction possible (mais conduisant à des complications inextricables) du volume des données .

Nous cherchons au contraire une technique de réduction s'appliquant à n'importe quel type de tableau, et conduisant à une reconstitution rapide du tableau de départ. Pour cela, nous nous placerons successivement dans les espaces vectoriels R^p et R^n (pour notre exemple : p = 300, n = 2 000).

I-2.1 - Ajustement dans R^p

Chacune des n lignes du tableau X est un vecteur, ou encore un point, de R^p. L'ensemble du tableau est un nuage de n points. Si par hasard ce nuage était contenu dans un sous-espace vectoriel à q dimensions de R^p, nous aurions résolu notre problème d'approximation.

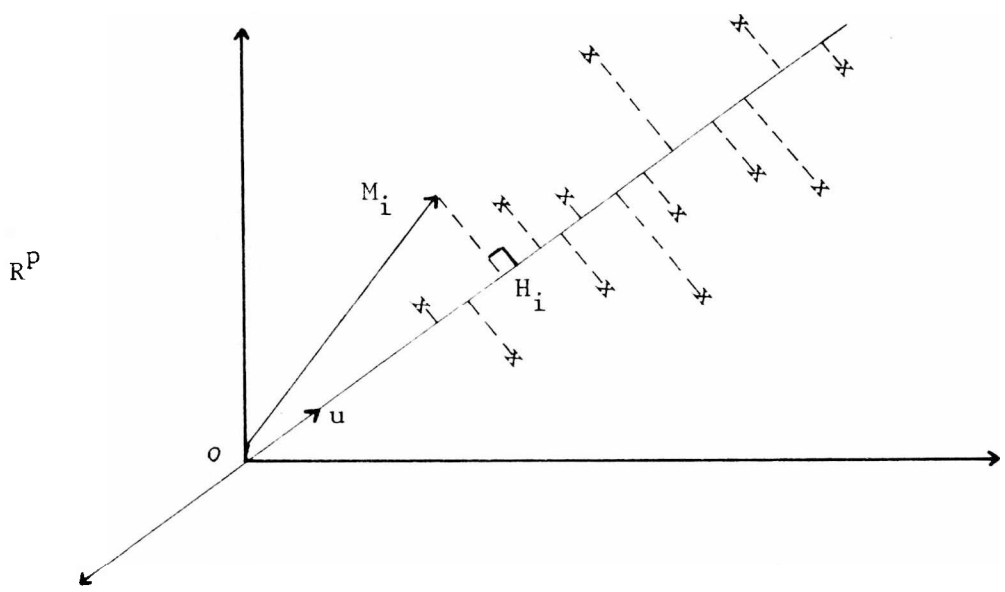
Ainsi, si les 2 000 points individus sont dans un sous-espace à 10 dimensions (ou plus généralement si leurs positions sont reconstituées de façon satisfaisante à partir de leurs positions dans ce sous-espace), il suffit de connaître la nouvelle base (10 vecteurs à 300 dimensions) et les nouvelles coordonnées des points dans cette

base (2 000 vecteurs à 10 dimensions) pour reconstituer les positions des points dans R^p, donc pour reconstituer les 600 000 coordonnées initiales. Or : 10 x 300 = 2 000 x 10 = 23 000.

On pourrait donc reconstituer 600 000 nombres à partir de 23 000.

Commençons donc à chercher un sous-espace vectoriel à une dimension qui ajuste au mieux le nuage de points.

Figure 1



Soit $\vec{u}$ un vecteur unitaire. On désignera par u la matrice-colonne associée et par $\tilde{u}$ sa transposée. Puisque u est unitaire, $\tilde{u}u = 1$.

La projection d'un vecteur $\vec{OM}$ sur le sous-espace à une dimension engendré par $\vec{u}$ n'est autre que le produit scalaire $\vec{OM}.\vec{u}$, somme des produits terme à terme des composantes de $\vec{OM}$ et de $\vec{u}$.

Chacune des n lignes du tableau X est un vecteur de R^p.

Le produit matriciel Xu est une matrice-colonne à n éléments, dont chaque élément est le produit scalaire d'une ligne de X par $\vec{u}$.

Ainsi, les n composantes de la matrice colonne Xu ne sont autres que les n projections OH_i des n points du nuage sur $\vec{u}$.

Un critère d'ajustement du sous-espace vectoriel consiste à minimiser la somme des carrés des écarts :

$$\sum_{i=1}^{n} \overline{M_i H_i}^2.$$

(C'est en fait le critère qui conduit aux calculs analytiques les plus simples).

D'après le théorème de PYTHAGORE : $\sum \overline{M_i H_i}^2 = \sum \overline{OM_i}^2 - \sum \overline{OH_i}^2$

Comme $\sum \overline{OM_i}^2$ est une quantité fixe, il revient au même de maximiser $\sum \overline{OH_i}^2 = [\widetilde{Xu}]\, Xu = \tilde{u}\, \tilde{X}\, Xu$.

Pour trouver u, on est donc conduit à chercher le maximum de la forme quadratique $\tilde{u}\, \tilde{X}\, Xu$, sous la contrainte $\tilde{u}u = 1$.

I-2.2 - Maximum d'une forme quadratique sous contrainte quadratique

A et B désignent deux matrices symétriques.

Soit à rendre maximum $\tilde{u}\, Au$, avec la contrainte $\tilde{u}\, Bu = 1$ (B est supposée inversible). Une règle de dérivation matricielle très simple à reconstituer nous dit que le vecteur des n dérivées de $\tilde{u}\, Au$ par rapport aux p composantes de u s'écrit :

$$\frac{\partial \tilde{u}\, Au}{\partial u} = 2\, Au$$

[Le vérifier en dérivant $\sum_{i,j} a_{ij} u_i u_j$ par rapport à u_i]

La recherche d'un maximum lié implique que s'annulent les dérivées du Lagrangien : $L = \tilde{u}\, Au - \lambda(\tilde{u}\, Bu - 1)$

Par suite $\frac{\partial L}{\partial u} = 2\, Au - 2\, \lambda Bu$

Ce qui prouve que : $\boxed{Au = \lambda Bu}$ (1)

Ou encore que : $B^{-1}Au = \lambda u$

Ainsi, u est vecteur propre de $[B^{-1}A]$.

En prémultipliant les deux membres de (1) par $\tilde{u}$, on obtient :

$\tilde{u}\, Au = \lambda$ (puisque $\tilde{u}\, Bu = 1$)

L'extremum cherché a pour valeur la valeur propre λ.

Comme nous cherchons un maximum, nous devrons retenir pour u le vecteur propre de $[B^{-1}A]$, correspondant à la plus grande valeur propre λ.

Dans le cas particulier qui nous intéresse, la matrice B est la matrice unité $[\delta_{ij}]$, notée parfois I.

Donc u est vecteur propre de A relatif à la plus grande valeur propre λ.

Nous avons choisi une norme $\tilde{u}\, Bu$ plus générale que la norme "euclidienne classique" $\tilde{u}u$ car les résultats de ce paragraphe nous serviront par la suite (Analyse des correspondances, discriminante). Pour fixer les idées, $\tilde{u}\, Bu$ peut représenter la norme usuelle lorsque les axes de départ ne sont pas orthogonaux.

Si nous cherchons un vecteur v, orthogonal à u dans ce système d'axe (donc tel que $\tilde{u}\, Bv = 0$) unitaire (donc tel que $\tilde{v}\, Bv = 1$), et qui rend maximum la forme quadratique $\tilde{v}\, Av$, nous sommes conduits à égaler à 0 les dérivées du Lagrangien.:

$$L = \tilde{v}\, Av - \lambda' \tilde{u}\, Bv - \lambda''(\tilde{v}\, Bv - 1)$$

Le vecteur des n dérivées de $\tilde{u}\, Bv$ par rapport à v vaut : Bu.

On a donc : $\dfrac{\partial L}{\partial v} = 2\, Av - \lambda' Bu - 2\, \lambda'' Bv = 0$

(λ' et λ'' sont deux multiplicateurs de Lagrange).

Montrons tout d'abord que $\lambda' = 0$

Prémultipliant $\frac{\delta L}{\delta v}$ par $\tilde{u}$, on obtient, puisque $\tilde{u}\,Bv = 0$ et

$\tilde{u}\,Bu = 1$: $2\,\tilde{u}\,Av = \lambda'$

Or la relation (1) transposée, nous apprend que $\tilde{u}\,A = \lambda\tilde{u}\,B$, d'où $\lambda' = 0$. Il reste la relation $Av = \lambda''Bv$.

Ceci prouve que v est également vecteur propre de $[B^{-1}A]$, et, comme précédemment, λ'' la valeur de l'extremum cherché, ce qui nous conduit à retenir pour v le vecteur propre de $B^{-1}A$ correspondant à la deuxième plus grande valeur propre.

La démonstration se généralise aisément aux q premiers vecteurs propres de $[B^{-1}A]$, tant que q n'excède pas le rang de A.

I-2.3 - Relation entre les ajustements dans R^p et R^n.

Le paragraphe II-2. nous montre que le vecteur unitaire u caractérisant le sous-espace à une dimension ajustant au mieux le nuage de points dans R^p est le vecteur propre de $[\tilde{X}X]$ correspondant à la plus grande valeur propre.

Le sous-espace à deux dimensions ajustant au mieux le nuage contient obligatoirement le sous-espace engendré par u. (Un raisonnement immédiat par l'absurde nous prouve que s'il ne contient pas u, il en existe un meilleur contenant u). On peut dont chercher un second vecteur de base de ce sous-espace, v, orthogonal à u, et maximisant $\tilde{v}\,\tilde{X}Xv$. Le paragraphe précédent nous montre également que v est le vecteur propre de $\tilde{X}X$ correspondant à la deuxième plus grande valeur propre. Plus généralement, le sous-espace à q dimensions qui ajuste au mieux (au sens des moindres carrés) le nuage de R^p est engendré par les q premiers vecteurs propres de la matrice symétrique $\tilde{X}X$.

Plaçons-nous maintenant dans l'espace R^n, où le tableau X peut être représenté par un nuage de p points, dont les n coordonnées constituent les colonnes dans X.

La recherche d'un vecteur unitaire $\vec{s}$ (puis d'un sous-espace à q dimensions, ajustant au mieux le nuage de R^n) nous conduiront à maximiser la somme des carrés des p projections sur $\vec{s}$, qui sont les composantes du vecteur $\tilde{X}s$.

La quantité à maximiser vaut : $[\tilde{X}s]\,\tilde{X}s = \tilde{s}\,X\tilde{X}s$, avec la contrainte $\tilde{s}s = 1$.

Nous serons, comme précédemment, amenés à retenir les q vecteurs propres de $X\tilde{X}$ correspondant aux q plus grandes valeurs propres.

Dans R^p, nous avions la relation $\tilde{X}\,Xu = \lambda u$ (2)

Dans R^n, nous avons $X\,\tilde{X}s = \mu s$ (3)

Prémultiplions les deux membres de (2) par X. On obtient :

$$X\,\tilde{X}\,Xu = \lambda Xu$$

Ceci nous prouve qu'à tout vecteur propre u de $\tilde{X}X$ relatif à une valeur propre λ non nulle correspond un vecteur propre Xu de $X\tilde{X}$, relatif à la même valeur propre.

Il est donc inutile de refaire les calculs de diagonalisation dans R^n, puisqu'une simple transformation linéaire (associée à la matrice X de départ) nous permet d'obtenir les vecteurs propres cherchés.

Un problème de norme reste à régler. Si u est unitaire, Xu a pour norme $\tilde{u}\,\tilde{X}Xu = \lambda$, le vecteur cherché s unitaire sera donc :

$$s = \frac{1}{\sqrt{\lambda}} Xu \qquad (4)$$

On a la relation symétrique : $u = \frac{1}{\sqrt{\lambda}}\tilde{X}s$ (5)

qui découle des relations (2) et (3).

I-2.4 - Reconstitution des données de départ.

Nous désignerons par u_i le i-ème vecteur propre de norme 1 de la matrice $\tilde{X}X$, supposée non singulière.

La relation (2) s'écrit $Xu_i = \sqrt{\lambda_i}\, s_i$ pour les i-ème axes de R^p et R^n.

Postmultiplions les deux membres de cette relation par $\tilde{u}_i$, et

sommons sur l'ensemble des axes:(Certains d'entre eux peuvent éventuellement correspondre à une valeur propre nulle. Ils complètent alors la base orthonormée formée par les autres).

$$X.\sum_i u_i \tilde{u}_i = \sum \sqrt{\lambda_i} \cdot s_i \tilde{u}_i$$

Or la matrice $U\tilde{U} = \sum_{i=1}^{p} u_i \tilde{u}_i$ n'est autre que la matrice unité d'ordre (p, p) car les p vecteurs propres "u_i" de $\tilde{X}X$ sont orthogonaux et de norme 1.

Si U désigne la matrice ayant en colonnes ces vecteurs propres, la matrice : $\tilde{U}\,U = I$ (matrice unité), d'où également $U\,\tilde{U} = I$.

On a donc la formule de reconstitution, les valeurs propres λ_i étant classés par ordre décroissant :

$$X = \sum_{i=1}^{p} \sqrt{\lambda_i} \cdot s_i \tilde{u}_i$$

Si les p - q plus petites valeurs propres sont très petites, on peut limiter la sommation aux q premiers termes :

$$X \simeq \sum_{i=1}^{q} \sqrt{\lambda_i}\, s_i \tilde{u}_i$$

Si $q \ll p$, on apprécie le gain réalisé en comparant les deux membres de cette relation : le vecteur $\sqrt{\lambda_i}\, s_i$ a n composantes et u_i a p composantes.

Les np termes de X sont approchés par les q(n + p) valeurs contenues dans le membre de droite.

I-3. ANALYSE GENERALE AVEC DES METRIQUES ET DES CRITERES QUELCONQUES

Trois étapes, lors de l'exposé du paragraphe précédent, peuvent donner lieu à des généralisations :

a) La construction même du nuage, qui peut résulter de transformations (simples de préférence) faites sur le tableau de valeur numérique X.

b) Le choix des distances entre les points, qui, au lieu d'être la somme des carrés des différences de coordonnées (distance euclidienne usuelle) peut être une distance euclidienne quelconque représentée par une forme quadratique associée à une matrice Q.

c) Le critère de maximisation, qui, au lieu d'être une simple somme de carré, peut être une forme quadratique, associée à une matrice S.

En général, il existera d'étroites relations entre les transformations initiales, les distances, les critères dans R^p et R^n.

Soient T_p, Q_p, S_p les matrices associées respectivement à la transformation, à la distance, au critère dans R^p, et T_n, Q_n, S_n leurs homologues dans R^n.

Si u désigne un vecteur unitaire de R^p, ($\tilde{u} Q_p u = 1$), le vecteur à n composantes des projections des n points (n lignes de $T_p X$) est maintenant le vecteur $T_p X Q_p u$, et sa norme selon S_p, qu'il faut rendre maximum, vaut : $\left[\tilde{u} Q_p \tilde{X} \tilde{T}_p\right] S_p \left[T_p X Q_p u\right]$.

L'équation aux vecteurs propres s'écrit, après simplification par Q_p :

$$\tilde{X} \tilde{T}_p S_p T_p X Q_p u = \lambda u$$

Heureusement, les six matrices précédentes sont toujours de forme très simple : en analyse des correspondances, elles sont toutes les six diagonales, et, lorsqu'elles ont même ordre, elles sont, soit égales, soit inverses l'une de l'autre.

Désignons par $D_p = T_p X$ le tableau transformé.

L'équation précédente s'écrit : $\tilde{D}_p S_p D_p Q_p u = \lambda u$

Le vecteur u s'appelle un axe factoriel de R^p, associé à λ.

La "forme linéaire" $v = Q_p u$ s'appelle facteur de R^p, associé à λ. (Ces deux notions sont confondues dans le cas du paragraphe précédent où $Q_p = I$).

Les projections des points sur l'axe u sont les composantes de $D_p Q_p u$. Ce sont ces valeurs qui intéressent l'utilisateur, puisqu'elles représentent l'approximation à une dimension du nuage.

Le facteur v vérifie l'équation : $Q_p \left[\tilde{D}_p S_p D_p\right] v = \lambda v$. La matrice entre crochets est la matrice d'inertie par rapport à l'origine du nuage de points.

I-4. PRATIQUE DE L'ANALYSE

Nous allons montrer, dans ce paragraphe, l'adaptation que doit subir l'analyse générale du § 3 lorsqu'il s'agit de décrire, et non plus seulement de réduire, des tableaux dont, rappelons-le, une au moins des dimensions est homogène. (Tableau du type : variables-individus).

Pour fixer les idées, le tableau X aura pour colonnes les 300 réponses ou mesures faites sur 2 000 individus. X est donc d'ordre 2 000, 300 ($n = 2\,000$, $p = 300$). Nous supposerons, afin de rester concrets, qu'il s'agit des valeurs de 300 types de dépenses annuelles relatives à 2 000 individus enquêtés.

Nous voulons avoir une idée de la structure de cet ensemble de 300 dépenses, ainsi que des similitudes éventuelles de comportement entre les individus enquêtés.

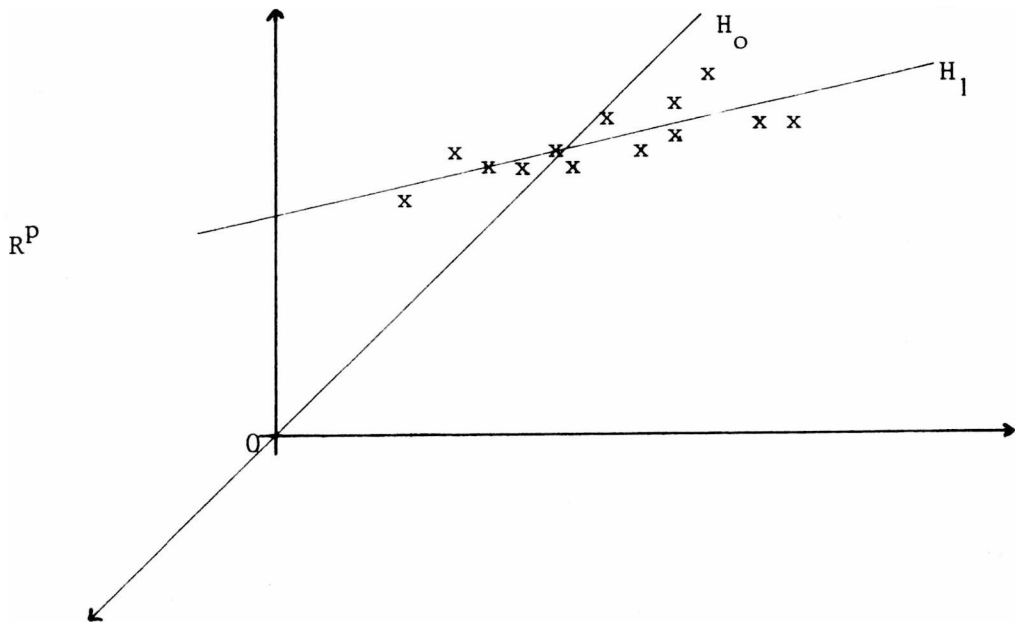
I-4.1 - Analyse dans R^p

Nous voulons, dans cet espace ajuster le nuage de n points par un sous-espace à une, puis deux dimensions, de façon à obtenir sur un graphique une représentation visuelle la plus fidèle possible des proximités existant entre les n individus vis-à-vis des p variables.

a) Ce n'est donc plus la somme des carrés des distances à l'origine projetées, qu'il faudra rendre maximum, mais la somme des carrés des distances entre tous les couples d'individus : autrement dit,

la droite d'ajustement H_1 ne doit pas être astreinte à passer par l'origine, comme H_0 (figure 2)

Figure 2



Si x_ℓ et x_k désignent les valeurs des projections de deux points individus ℓ et k sur H_1, on a la relation :

($\sum_{k,\ell}^{n}$ désigne une sommation pour tout k et tout ℓ inférieurs ou égaux à n).

$$\sum_{k,\ell}^{n} (x_\ell - x_k)^2 = n\sum_{\ell} x_\ell^2 + n\sum_{k} x_k^2 - 2\left(\sum_{\ell} x_\ell\right)\left(\sum_{k} x_k\right)$$

$$= 2n^2 \left(\frac{1}{n}\sum_{\ell} x_\ell^2 - \bar{x}^2\right) = 2n \sum_{\ell} (x_\ell - \bar{x})^2$$

$\bar{x}$ désigne la moyenne des projections des n individus, donc également la projection du point moyen G du nuage sur H_1.

Ainsi, si l'origine est prise en G, la quantité à maximiser sera à nouveau la somme des carrés des distances à l'origine.

Le sous-espace cherché correspondra à l'analyse générale du tableau transformé D, de terme général $d_{ij} = x_{ij} - \bar{x}_j$.

b) La distance entre deux individus ℓ et k s'écrit, dans R^p :

$$d^2(\ell,k) = \sum_{j=1}^{p} (x_{\ell j} - x_{kj})^2$$

Dans cette somme, il peut exister des valeurs de j pour lesquelles les variables correspondantes sont d'échelles très diverses : (exemple : j = 3 : dépense de timbres poste, j = 142 : dépenses de loyer). On peut parfois, surtout lorsque les unités de mesures ne sont pas les mêmes, désirer faire jouer à chaque variable un rôle identique dans la définition des proximités entre individus.

On corrige alors les échelles en adoptant la distance :

$$d^2(\ell,k) = \sum_{j=1}^{p} \left(\frac{x_{\ell j} - x_{kj}}{s_j} \right)^2$$

s_j désignant l'écart-type de la variable j.

Finalement, nous retiendrons que l'ACP dans R^p du tableau brut X est l'analyse générale de D, de terme général :

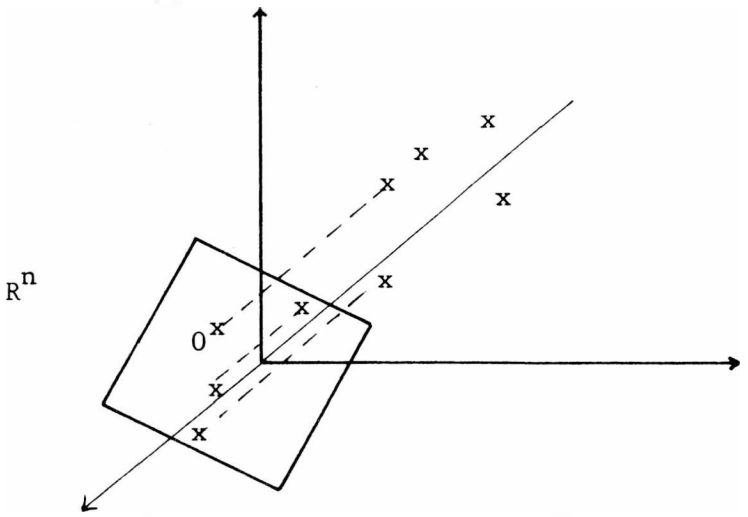
$$d_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

I-4.2 - Analyse dans R^n.

Dans cet espace, les points représentent des variables. Les transformations que nous allons être amenés à faire conduiront à la même formule analytique de changement de variable :

$$d_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Cependant, dans cet espace, cette même formule correspond à des opérations géométriques tout à fait différentes.



Ainsi, le point représentant la variable "dépense de timbres poste" sera situé près de l'origine (valeurs faibles pour les n individus), alors que le point "dépense de loyer" sera au contraire loin de l'origine, dans la direction de la première bissectrice.

Or ces caractéristiques de <u>niveau</u> ne nous intéressent pas ici. (Il nous suffit de consulter la liste des dépenses moyennes pour être renseignés).

Nous supprimerons donc cet axe de dispersion trivial en projetant le nuage sur un sous-espace à n-1 dimensions orthogonal à cet axe.

La matrice P associée à cette projection a pour terme général :

$$P_{ij} = \delta_{ij} - \frac{1}{n}$$

Ou encore, en notant I la matrice unité, U la matrice dont tous les éléments valent 1,

$$P = I - \frac{1}{n} U$$

Comme on peut le vérifier à titre d'exercice, cette matrice transforme un vecteur x, de composantes $x_1, x_2, \ldots\ldots, x_n$ en un vecteur de composantes :

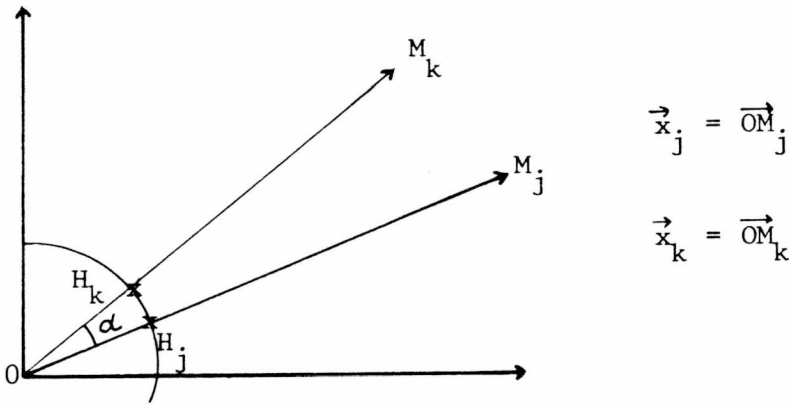
$$x_1 - \bar{x},\ x_2 - \bar{x},\ \ldots\ldots\ x_n - \bar{x} \quad \left(\text{avec } \bar{x} = \frac{1}{n} \sum_{1}^{n} x_1\right)$$

Une fois projeté, le nuage de points reste encore trop fruste :

la proximité de deux points reste difficile à interpréter ; en effet, la distance d'un point j à l'origine s'écrit :

$$\left|\vec{x}_j\right|^2 = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

Figure 3



Ce n'est autre, au coefficient $1/n$ près, que la variance de la variable j. Or nous savons surtout interpréter les proximités en termes de corrélation, et le coefficient de corrélation entre les variables j et k n'est autre que

$$\rho_{kj} = \cos\alpha = \frac{\vec{x}_j \cdot \vec{x}_k}{\sqrt{|\vec{x}_j| \cdot |\vec{x}_k|}}$$

Nous nous intéresserons donc surtout aux vecteurs unitaires $\overrightarrow{OH}_k$, $\overrightarrow{OH}_j$ portés par les vecteurs $\vec{x}_k$ et $\vec{x}_j$. Les nouvelles distances s'interprétent alors aisément. Dans le triangle OH_kH_j, on a en effet la relation :

$$H_kH_j^2 = OH_k^2 + OH_j^2 - 2\, OH_k . OH_j \cos\alpha$$

$$= 1 + 1 - 2\cos\alpha = 2(1 - \rho_{kj})$$

Ainsi, dans notre nouveau nuage, une distance nulle signifie que $\rho_{kj} = 1$ donc que les deux variables concernées sont liées par une relation linéaire directe. Le carré de distance vaut 2 dans le cas d'indépendance ($\rho_{ij} = 0$) et 4 dans le cas de liaison inverse ($\rho_{ij} = -1$).

Soit : $C\,\tilde{D}v = \frac{\lambda}{n}\,\tilde{D}v$

Les vecteurs u et $\tilde{D}v$ sont donc colinéaires. Comme u est unitaire et comme $\tilde{D}v$ a pour longueur : $\tilde{v}\,D\,\tilde{D}v = \lambda$, on a

$$\tilde{D}v = \sqrt{\lambda}\,u$$ (résultat déjà établi lors de l'analyse générale)

Réciproquement : $Du = \sqrt{\lambda}\,v$

Ainsi : les coordonnées des points sur un axe factoriel dans un espace sont proportionnelles aux cosinus directeurs de l'axe factoriel correspondant dans l'autre espace.

Notons que les abscisses et les ordonnées sur le graphique des points-variables ont une interprétation simple :

La matrice D a pour terme général $d_{ij} = \dfrac{x_{ij} - \bar{x}_j}{s_j}$

Les éléments de chaque colonne ont donc pour variance 1, et pour moyenne 0.

Le vecteur Du_i a pour composantes les n projections des points-individus sur l'axe i. La moyenne des valeurs de ces composantes est nulle (comme combinaisons linéaires de variables de moyennes nulles). La variance de ces valeurs vaut :

$$\frac{1}{n}\,\tilde{u}_i\,\tilde{D}\,Du_i = \frac{\lambda_i}{n}\,.$$

Par suite, le coefficient de corrélation de la variable j avec les coordonnées des n points sur l'axe i sera la j-ème composante du vecteur :

$$\frac{1}{\sqrt{n\lambda_i}}\,\tilde{D}\,Du_i = \frac{1}{\sqrt{n\lambda_i}}\,.\,\lambda u_i = \sqrt{\frac{\lambda_i}{n}}\,.\,u_i$$

Pour pouvoir interpréter les positions respectives des variables et des axes en termes de corrélation, nous procéderons donc au changement d'échelle qui consiste à multiplier par $\sqrt{\frac{\lambda_i}{n}}$ les coordonnées des variables sur ces axes : ceci permet de lire directement la contribution d'une variable à un axe.

Finalement, dans R^n, nous sommes conduits à faire l'analyse du tableau transformé :

$$d_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

selon la même formule que précédemment. (Formule non symétrique par rapport aux deux indices i et j).

Les transformations dans les deux espaces conduisent au même tableau D. Par suite, l'analyse en composantes principales est strictement un cas particulier du modèle général développé au §2.

I-4.3 - Etapes du calcul.

Le vecteur unitaire u de R^p engendrant le sous-espace d'ajustement vérifie l'équation :

$$\tilde{D}\, D u = \lambda u \qquad \text{rappelons que } d_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Or la matrice $C = \frac{1}{n} \tilde{D}\, D$ n'est autre que la matrice des corrélations expérimentales des p variables (puisque $c_{jj'} = \frac{1}{n} \sum_{i=1}^{n} d_{ij}\, d_{ij'}$)

Les facteurs vérifient donc la relation : $Cu\left(\frac{\lambda}{n}\right) u$
Ce sont les vecteurs propres de la matrice des corrélations des variables.

Les projections des individus sur cet axe sont les composantes de Du. Si u_1 et u_2 désignent les deux premiers facteurs, les abscisses et les ordonnées des points sur le graphique seront respectivement les composantes de Du_1 et Du_2.

Le vecteur unitaire v de R^n engendrant le sous-espace d'ajustement vérifie :

$$D\, \tilde{D} v = \lambda v$$

Et les projections des p points variables sur les deux premiers axes factoriels de R^n sont, de même, les composantes de $\tilde{D}v_1$ et $\tilde{D}v_2$.

Prémultipliant les deux membres et la relation précédente par $\frac{1}{n}\, \tilde{D}$, il vient :

$$\frac{1}{n}\, \tilde{D}\, D\, \tilde{D} v = \frac{\lambda}{n} \tilde{D} v$$

Les étapes de calcul sont donc les suivantes, dans le cas où l'on désire se limiter à deux dimensions :

1° - diagonalisation de la matrice des corrélations

2° - construction du graphique des points-variables à partir des composantes de

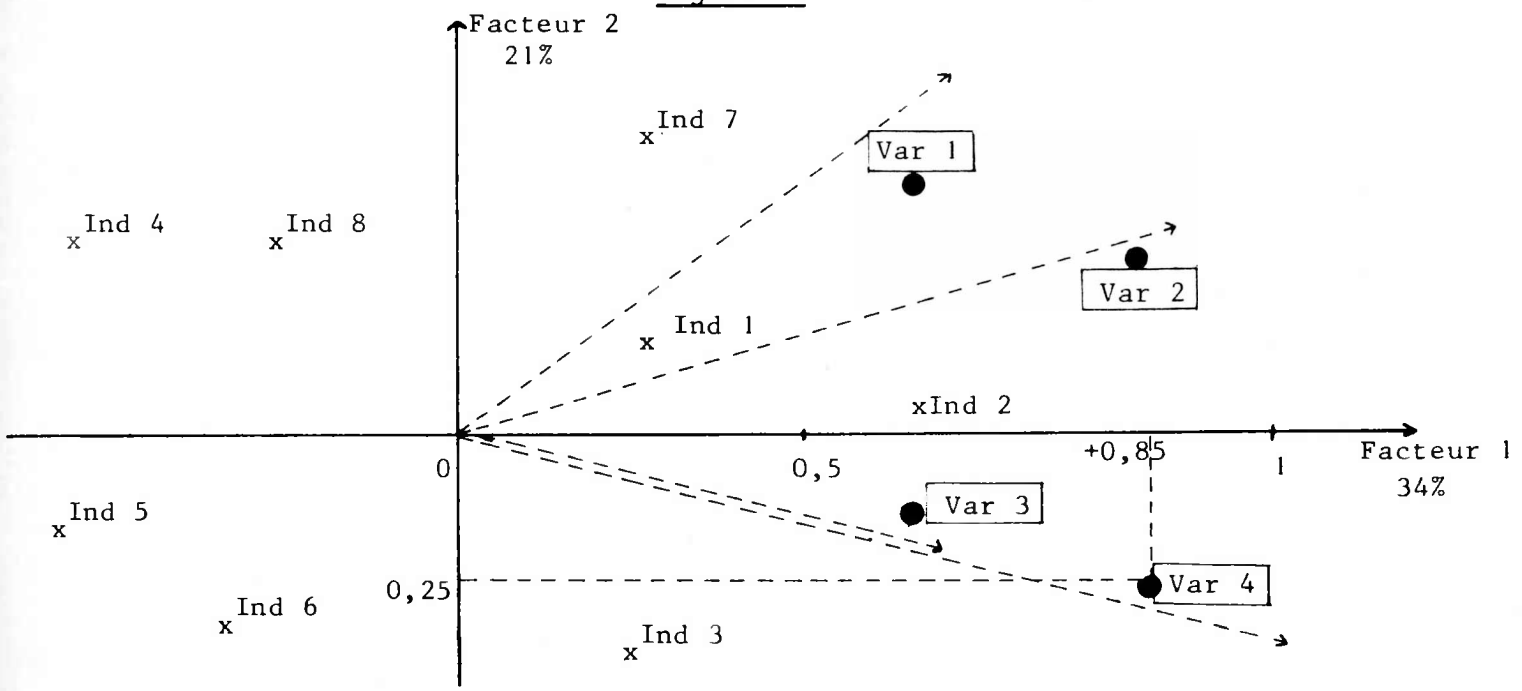
$$u_1 \sqrt{\frac{\lambda_1}{n}} \quad \text{et} \quad u_2 \sqrt{\frac{\lambda_2}{n}}$$

3° - construction du graphique des points-individus, à partir des composantes de Du_1 et Du_2.

I-4.4 - Présentation des résultats et règles d'interprétation.

Il est commode de représenter les figures obtenues dans chacun des deux espaces sur un même graphique, à condition de respecter certaines règles d'interprétation qui découlent directement des considérations théoriques précédentes. L'idéal serait en fait de représenter un des deux ensembles sur un calque. En pratique, on représentera les deux ensembles par des caractères très différents.

Figure 4



La figure 4 nous donne un exemple d'édition de résultats pour $p = 4$ et $n = 8$ (quatre variables et huit individus).

Comme on peut le constater, les individus se répartissent de façon équilibrée autour de l'origine, qui est, rappelons-le, leur centre de gravité. Par contre, les variables peuvent très bien être situées toutes d'un même côté de l'un des axes. (Car l'analyse du nuage des p points-variables dans R^n se fait à partir de l'origine). Rappelons que le cosinus de l'angle sous lequel on voit deux points-variables dans R^n n'est autre que le coefficient de corrélation entre ces deux variables. Selon le pouvoir explicatif des facteurs (définis plus loin) cette propriété sera plus ou moins bien conservée par projection.

Si les proximités entre individus s'interprètent en termes de "similitude de comportement vis-à-vis des variables", les proximités entre variables en termes de corrélation, il faut bien se garder d'interpréter la distance séparant un point-variable d'un point-individu, car ces deux points ne font pas partie d'un même nuage dans un même espace.

Cependant, il est licite de comparer les positions respectives de deux individus vis-à-vis de l'ensemble des variables, ou de deux variables vis-à-vis de l'ensemble des individus. En effet, les axes en pointillé ne sont autres, à une dilatation près, que les projections des axes de R^p (correspondant chacun à une variable) sur le plan des deux premiers facteurs, qui ajuste au mieux dans cet espace le nuage des n points-individus. (Puisque les coordonnées des points-variables sont précisément proportionnelles aux composantes des cosinus directeurs des nouveaux axes : soit $\vec{u}$, unitaire, porté par l'axe 1; le produit scalaire de $\vec{u}$ par $\vec{e}_i$, (dont la j-ème composante vaut δ_{ij}) vaut :

$$\langle \vec{u}, \vec{e}_i \rangle = \sum_{j=1}^{p} \delta_{ij} u_j = u_i$$

La coordonnée de la i-ème variable sur cet axe est précisément :

$$u_i \sqrt{\frac{\lambda_1}{n}}).$$

La figure 4 nous fournit donc une perspective caricaturale du système d'axe original, tenant compte des liaisons existant entre les variables initiales.

On peut donc dire, par exemple, que les individus 4 et 5 ont des comportements relativements similaires, caractérisés par des valeurs faibles pour les 4 variables, contrairement à l'individu 2, qui a au contraire le meilleur "score" pour ces variables.

Ainsi, avec les précautions d'interprétation signalées plus haut, la représentation simultanée des deux ensembles permet de comprendre les proximités à l'intérieur d'un même ensemble, en exhibant les variables responsables de ces proximités.

Les pourcentages qui figurent sur les axes définissent les "pouvoirs explicatifs" des facteurs : ils représentent la part de la variance (ou inertie) totale expliquée par chaque facteur. Ainsi, le pourcentage de 34% associé au premier facteur signifie que la première valeur propre représente 34% de la somme des valeurs propres (trace de la matrice à diagonaliser).

Comme nous le signalerons encore à propos de l'analyse des correspondances, il s'agit d'une mesure exagérément pessimiste du pouvoir explicatif des facteurs, liée parfois de façon assez arbitraire, au codage des données.

II - ANALYSE DES CORRESPONDANCES (A.C.)

L'analyse des correspondances, technique due au Professeur BENZECRI, a un domaine d'application différent de l'analyse en composantes principales. Alors que l'on réserve cette dernière aux tableaux de mesures éventuellement hétérogènes, l'analyse des correspondances fournit des descriptions optimales des tableaux de contingence (ou de dépendance, ou encore tableaux croisés), et, par extension, des descriptions satisfaisantes des tableaux de codages discontinus.

Comme dans le cas de l'analyse en composantes principales, nous allons être amenés à effectuer des transformations du tableau de données, afin d'éliminer l'information exogène, c'est-à-dire ici "connue a priori".

Nous raisonnerons sur un exemple concret : le tableau de contingence croisant les communes de la région parisienne (373 communes dans l'agglomération) et 29 catégories d'activité de la population. Nous poserons $p = 29$ et $n = 373$.

Nous allons voir comment la construction du nuage, le choix de la distance, le choix du critère d'ajustement s'imposent de par la nature même des données analysées. Nous donnerons plus loin d'autres présentations de l'analyse des correspondances.

II-1. GEOMETRIE DES NUAGES ET CRITERES D'AJUSTEMENT

II-1.1 - Construction des nuages

Plaçons-nous tout d'abord dans l'espace R^p, où le tableau est représenté par 373 points-communes.

Le tableau de données sera toujours représenté par la matrice X d'ordre (n, p) .

- x_{ij} représente le nombre d'individus habitant dans la commune "i" et appartenant à la catégorie socio-professionnelle "j".

Si les composantes des vecteurs de R^p sont constituées par des effectifs bruts tels que x_{ij}, les proximités existant entre les points-communes seront peu intéressantes à interpréter. Il existe en effet des communes très peuplées, pour lesquelles les 29 composantes seront élevées, et des communes très petites pour lesquelles ces composantes sont toutes faibles, et qui seront situées près de l'origine. En fait, la taille des communes est souvent une conséquence d'un découpage arbitraire.

Ce ne sont pas les effectifs bruts qui nous intéressent, mais ce que l'on appelle les "profils socio-professionnels" des communes, c'est-à-dire la proportion de chacune des C.S.P. dans cette commune.

De la même façon, dans l'espace R^n, ce ne sont pas les effectifs bruts des points - C.S.P., mais les "profils géographiques ou communaux" de chacune des C.S.P. qui vont retenir notre attention.

Notation - Nous désignerons par $x_{i.} = \sum_{j=1}^{p} x_{ij}$ l'effectif total de la population de la commune i;

par $x_{.j} = \sum_{i=1}^{n} x_{ij}$ l'effectif total de la C.S.P. j ;

par $x.. = \sum x_{ij}$ l'effectif total de la population concernée.

Nous prendrons comme composante du i-ème vecteur de R^p :

$$\left\{ \frac{x_{ij}}{x_{i.}} \right\} \quad j = 1, 2, \ldots, p$$

Nous prendrons comme composante du j-ème vecteur de R^n :

$$\left\{ \frac{x_{ij}}{x_{.j}} \right\} \quad i = 1, 2, \ldots, n$$

Ici, nous devrons noter une première différence fondamentale avec l'analyse en composantes principales : les transformations faites sur les données brutes dans les deux espaces sont analogues (car les ensembles mis en correspondance jouent des rôles symétriques). Elles correspondent à des transformations analytiques

différentes : le tableau des nouvelles coordonnées dans R^p n'est pas le simple transposé de celui des nouvelles coordonnées dans R^n. (Alors qu'en A.C.P., des transformations très différentes conduisaient à une même formule analytique).

Nous noterons les fréquences calculables à partir du tableau X de la façon suivante :

$$f_{ij} = \frac{x_{ij}}{x..}$$

et de façon analogue :

$$f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{x_{i.}}{x..}$$

$$f_{.j} = \sum_{i=1}^{n} f_{ij} = \frac{x_{.j}}{x..}$$

On a la relation suivante :

$$\frac{f_{ij}}{f_{.j}} = \frac{x_{ij}}{x_{.j}} \quad \text{et} \quad \frac{f_{ij}}{f_{i.}} = \frac{x_{ij}}{x_{i.}} \quad \text{pour tout } i \text{ et } j.$$

Nous raisonnerons dorénavant en termes de fréquences.

II-1.2 - Choix des distances

Le fait d'avoir choisi des "profils" (qui sont des fréquences conditionnelles) pour construire les nuages de points, et le souci d'observer une certaine invariance des résultats va nous conduire à adopter une distance différente de la distance euclidienne usuelle.

La distance entre deux communes i et i' sera donnée par la formule :

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

De façon symétrique, la distance entre deux C.S.P. j et j' vaut :

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

Ces distances ne diffèrent en fait de la métrique euclidienne usuelle que par la pondération de chaque carré par les inverses des fréquences correspondant à chaque terme.

Pourquoi avoir choisi cette distance ?

Parce qu'elle vérifie la propriété d'équivalence distributionnelle, qui s'exprime comme suit :

1) si l'on agrège deux catégories socio-professionnelles ayant des profils identiques, alors les distances entre communes seront inchangées.

2) si l'on agrège deux communes ayant des profils socio-professionnels identiques, les distances entre C.S.P. seront inchangées.

Cette propriété est fondamentale, car elle garantit une certaine invariance des résultats vis-à-vis de la nomenclature choisie pour la construction des classes.

D'un point de vue strictement technique, il est logique que deux points confondus dans R^p ou dans R^n, puissent être considérés comme un seul point correspondant aux effectifs des deux classes réunies.

La typologie des communes sera ainsi peu bouleversée si l'on agrège les catégories socio-professionnelles voisines quant à leurs répartitions géographiques.

Ainsi, on ne perd pas d'informations en agrégeant certaines classes, et l'on n'en gagne pas en subdivisant indéfiniment des classes homogènes. On peut ainsi vérifier que la typologie des C.S.P. obtenue à partir des 80 quartiers de Paris n'apporte pas d'information très nouvelle par rapport à celle obtenue à partir des 20 arrondissements.

Démontrons cette propriété lorsqu'il s'agit d'agréger deux communes i_1 et i_2 en une commune i_o, dont la fréquence relative de population, $f_{i_o.}$, vérifie :

$$f_{i_o.} = f_{i_1.} + f_{i_2.} \qquad (1)$$

Dans l'expression de la distance $d^2(j,j')$ entre deux C.S.P. j et j', seuls deux termes T_1 et T_2 font intervenir i_1 et i_2

$$T_1 + T_2 = \frac{1}{f_{i_1.}}\left(\frac{f_{i_1 j}}{f_{.j}} - \frac{f_{i_1 j'}}{f_{.j'}}\right)^2 + \frac{1}{f_{i_2.}}\left(\frac{f_{i_2 j'}}{f_{.j}} - \frac{f_{i_2 j'}}{f_{.j'}}\right)^2$$

Ils sont remplacés, après agrégation par T_o tel que :

$$T_o = \frac{1}{f_{i_o.}}\left(\frac{f_{i_o j}}{f_{.j}} - \frac{f_{i_o j'}}{f_{.j'}}\right)^2$$

Il s'agit de montrer que $T_o = T_1 + T_2$ (2)

Or, T_o s'écrit :

$$T_o = f_{i_o.}\left(\frac{f_{i_o j}}{f_{i_o.} f_{.j}} - \frac{f_{i_o j'}}{f_{i_o.} f_{.j'}}\right)^2$$

T_1 et T_2 s'écrivent de façon identique ; les trois quantités entre parenthèses sont égales, puisque les profils de i_1, i_2 et i_o sont identiques.

La relation (2) découle alors de (1)

II-1.3 - Choix du critère d'ajustement

Dans la construction des nuages de R^p et de R^n, le choix des profils comme coordonnées donne à toutes les communes et les C.S.P. la même importance.

Il est naturel de munir, pour le calcul de l'ajustement, chaque

point d'une masse proportionnelle à sa fréquence, afin de ne pas privilégier les classes d'effectifs faibles, et de respecter la répartition réelle de la population.

Nous ne chercherons donc plus les extrema de quantités du type $\sum_k Z_k^2$ mais du type $\sum_k f_k . Z_k^2$ ou $\sum_\ell f_\ell . Z_\ell^2$ selon l'espace choisi.

II-1.4 - Récapitulation.

Espace : R^p	Espace : R^n
n points $i = 1, \ldots n$	p points $j = 1, \ldots p$
p coordonnées du point i : $\left\{\frac{f_{ij}}{f_{i.}}\right\}$ $j = 1, \ldots p$	n coordonnées du point j : $\left\{\frac{f_{ij}}{f_{.j}}\right\}$ $i = 1, \ldots n$
Masse du point i : $f_{i.}$	Masse du point j : $f_{.j}$
Distance de deux points i et i' $d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$	Distance de deux points j et j' $d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2$

Pour faire le lien avec l'analyse générale des § 2 et 3 précédents, il est utile d'utiliser des notations matricielles qui s'avèrent cependant extrêmement lourdes, car les matrices définissant les transformations initiales, les distances, les critères, sont toutes diagonales, et peuvent donc être caractérisées par la variation d'un seul indice.

X désignant toujours la matrice de départ d'ordre (n, p), on notera F le tableau de fréquence associé :

$$F = \frac{1}{x..} X \qquad (\text{avec } x.. = \sum_{ij} x_{ij})$$

Nous désignerons par S la matrice d'ordre (p,p) diagonale, dont le j-ème élément diagonal vaut $s_{jj} = f_{.j}$ (ou $s_{ij} = \delta_{ij} f_{.j}$), par T la matrice d'ordre (n,n) diagonale, dont le i-ème élément diagonal vaut $t_{ii} = f_{i.}$ (ou $t_{ij} = \delta_{ij} f_{i.}$).

Ces deux matrices et le tableau F vont nous permettre de récapituler à nouveau les éléments de départ de l'analyse :

Espace : R^p	Espace : R^n
Les n points correspondent aux n lignes de $\underset{nn}{T^{-1}} \underset{np}{F}$	Les p points correspondent aux p colonnes de $\underset{np}{F} \underset{pp}{S^{-1}}$ (ou p lignes de $S^{-1} \tilde{F}$)
Critère d'ajustement : Forme quadratique dont la matrice est T	Critère : Forme quadratique associée à S
Distance : F.Q. associée à S^{-1}	Distance : F.Q. associée à T^{-1}

II-2. CALCUL DES AXES FACTORIELS ET DES FACTEURS

Il existe une symétrie totale entre les indices i et j. Il nous suffira de nous limiter à un espace, R^p par exemple, les démonstrations dans l'autre espace s'en déduisant par permutation des indices i et j, c'est-à-dire transposition de F et substitution réciproque des matrices S et T.

Dans les deux espaces, nous voulons représenter graphiquement les proximités entre profils. Nous nous placerons donc, dans chaque espace, aux centres de gravité des nuages.

Cependant, et c'est là une des particularités de l'analyse des correspondances, il est équivalent de procéder à l'analyse par rapport à l'origine ou par rapport aux centres de gravité, à condition de négliger (dans le premier cas) le premier axe factoriel qui joint l'origine au centre de gravité.

Nous commencerons donc à effectuer des analyses générales, puis nous montrerons ces équivalences.

II-2.1 - Analyse dans R^p, calcul des facteurs

Les n points sont les n lignes de $T^{-1}F$.

Soit u un vecteur unitaire, tel que $\tilde{u}S^{-1}u = 1$

Le vecteur des n projections s'écrit : $T^{-1}FS^{-1}u = v$
La norme de ce vecteur, selon le critère T vaut $\tilde{v}Tv$

soit : $$\tilde{u}\, S^{-1}\, \tilde{F}\, T^{-1} F\, S^{-1} u$$

Il faut rendre maximum cette quantité, avec la contrainte $\tilde{u}\, S^{-1} u = 1$. Ce problème a déjà été rencontré au § 3 de l'A.C.P. u est vecteur propre de

$$Q = \tilde{F}\, T^{-1} F\, S^{-1}$$

correspondant à la plus grande valeur propre λ: $Qu = \lambda u$.

Le terme général de Q, $q_{jj'}$, s'écrit :

$$q_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j'}}$$

Q n'est pas symétrique, mais nous verrons que l'on peut se ramener à la recherche de vecteurs propres et de valeurs propres d'une matrice symétrique.

Le vecteur $\varphi = S^{-1}u$ est appelé le premier facteur.

Les projections des n points sont les composantes de

$$T^{-1}FS^{-1}u = T^{-1}F\varphi.$$

II-2.2 - Liaison avec l'analyse dans R^n

De la même façon, dans R^n, on doit rendre maximum :

$$\tilde{w}\, T^{-1}\, F\, S^{-1}\, \tilde{F}\, T^{-1} w \quad \text{avec} \quad \tilde{w}\, T^{-1} w = 1$$

L'axe factoriel w est donc le vecteur propre de

$$R = F\, S^{-1}\, F\, T^{-1}$$

correspondant à la plus grande valeur propre.

Réécrivons l'équation, $Qu = \lambda u$:

$$\tilde{F}\, T^{-1}\, F\, S^{-1} u = \lambda u$$

Prémultiplions les deux membres par $F\, S^{-1}$:

$$F\, S^{-1}\, \tilde{F}\, T^{-1} (F\, S^{-1} u) = \lambda (F\, S^{-1} u)$$

Ainsi, comme dans le cas de la métrique usuelle, w est proportionnel à $F\, S^{-1} u$. Comme la T^{-1}- norme de $F\, S^{-1} u$ vaut λ, et que $\tilde{w}\, T^{-1} w = 1$, on doit poser :

$$w = \frac{1}{\sqrt{\lambda}}\, F\, S^{-1} u$$

(Symétriquement : $u = \frac{1}{\sqrt{\lambda}}\, \tilde{F}\, T^{-1} w$)

Les coordonnées des p points-variables sur cet axe sont les composantes de :

$$z = S^{-1}\, \tilde{F}\, T^{-1} w = S^{-1}\, \tilde{F}\, \psi \qquad (\text{où} \quad \psi = T^{-1} w)$$

Comme précédemment, ψ est appelé facteur, correspondant à la valeur propre λ.

- Remarquons que : $\psi = T^{-1} w = \frac{1}{\sqrt{\lambda}}\, T^{-1}\, F\, S^{-1} u = \frac{1}{\sqrt{\lambda}}\, T^{-1}\, F\, \varphi$ (3)

De même : $\varphi = S^{-1} u = \frac{1}{\sqrt{\lambda}}\, S^{-1}\, \tilde{F}\, T^{-1} w = \frac{1}{\sqrt{\lambda}}\, S^{-1}\, \tilde{F}\, \psi$ (4)

- Ces deux derniers ensembles de relations nous montrent que les coordonnées des points sur un axe factoriel dans un espace sont proportionnelles aux composantes des facteurs, de l'autre espace, correspondant aux mêmes valeurs propres.

Les relations (3) et (4) s'écrivent directement :

$$\psi_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{.j}}\, \varphi_j \qquad (3')$$

$$\varphi_j = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{n} \frac{f_{ij}}{f_{.j}} \psi_j \qquad (4')$$

Ainsi, au coefficient $\frac{1}{\sqrt{\lambda}}$ près, les points représentatifs d'un nuage sont les barycentres (car, par exemple : $\sum_j \frac{f_{ij}}{f_{i.}} = 1$) des points représentatifs de l'autre nuage.

II-2.3 - Remarques et mise en oeuvre du calcul

a) Analyse par rapport aux centres de gravité

Nous raisonnerons pour fixer les idées dans R^p.

Chaque point i a pour coordonnée $\left\{\frac{f_{ij}}{f_{i.}}\right\}_{j = 1,....p}$, et est muni de la masse $f_{i.}$.

Le centre de gravité du nuage, ou point moyen, a pour j-ème composante :

$$G_j = \sum_{i=1}^{n} f_{i.} \left(\frac{f_{ij}}{f_{i.}}\right) = \sum_{i=1}^{n} f_{ij} = f_{.j}$$

L'analyse par rapport au centre de gravité revient à remplacer :

$$\frac{f_{ij}}{f_{i.}} \text{ par } \left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right) \text{ ou encore } \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.}}$$

- Remarquons que le nuage est dans le sous-espace (affin) à $p-1$ dimensions H défini par la relation $\sum_j (f_{ij}/f_{i.}) = 1$, pour tout i.

Le centre de gravité appartient bien entendu à ce sous-espace, ainsi que tous les axes factoriels décrivant réellement les positions des points.

Tout vecteur libre de ce sous-espace est tel que la somme de

ses composantes est nulle. En effet, si x et y sont deux points de H dont les coordonnées sont x_j $(j = 1, \ldots, p)$ et y_j $(j = 1, \ldots, p)$ on a la relation :

$$\sum_{j=1}^{p} x_j = \sum_{j=1}^{p} y_j = 1 \quad \text{d'où :} \quad \sum_{j=1}^{p} (x_j - y_j) = 0$$

En particulier, les axes factoriels auront chacun des composantes de somme nulle.

La matrice Q, du § 2.1, devient, après substitution de $\left(\frac{f_{ij}}{f_{i.}}\right)$ par $\left(\frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}}\right)$:

$$Q' = (q'_{jj'}) = \left(\sum_{i=1}^{n} \frac{(f_{ij} - f_{i.}f_{.j})(f_{ij'} - f_{i.}f_{.j'})}{f_{i.}f_{.j'}}\right)$$

En développant $q_{jj'}$, et en tenant compte des relations :

$$\sum_i f_{i.} = \sum_j f_{.j} = \sum_i \sum_j f_{ij} = 1$$

$$\sum_i f_{ij} = f_{.j}$$

Il vient : $q'_{jj'} = \sum_i \frac{f_{ij}f_{ij'}}{f_{i.}f_{.j'}} - f_{ij} = q_{jj'} - f_{.j}$

Si u est un axe factoriel issu de G, il vérifie :

$$\sum_{j'=1}^{p} q'_{jj'}u_{j'} = \lambda u_j$$

Soit $$\sum_{j'} q_{jj'}u_{j'} - f_{.j}\sum_{j'} u_{j'} = \lambda u_j$$

Comme $\sum u_{j'} = 0$, puisque $u \in J$

$$\sum_{j'} q_{jj'} u_{j'} = \lambda u_j$$

Par suite, tout vecteur propre de Q' <u>appartenant à H</u> est vecteur propre de Q, relatif à la même valeur propre.

(La droite joignant l'origine à G est, comme on peut le vérifier aisément, vecteur propre de Q relatif à la valeur propre 1, et vecteur propre de Q' relatif à la valeur propre 0).

Il nous suffit donc de diagonaliser Q, en négligeant la valeur propre égale à 1.

b) <u>"Symétrisation" de Q.</u>

La matrice $(q_{jj'}) = \left(\sum_{i=1}^{n} \frac{f_{ij}\, f_{ij'}}{f_{i.}\, f_{.j'}} \right)$ n'est pas symétrique.

On l'écrit également : (cf. II.1)

$$Q = \tilde{F}\, T^{-1}\, F\, S^{-1}$$

La matrice $A = \tilde{F}\, T^{-1}\, F$ est symétrique. La matrice S^{-1} est diagonale.

Remarquons que $S^{-1} = \left(\frac{\delta_{ij}}{f_{j.}} \right) = S^{-\frac{1}{2}}\, S^{-\frac{1}{2}}$ ou $s_{ij}^{-\frac{1}{2}} = \frac{\delta_{ij}}{\sqrt{f_{j.}}}$

D'où $$Q = A\, S^{-\frac{1}{2}}\, S^{-\frac{1}{2}}$$

On a la relation $Q\, u = \lambda u$, soit : $A\, S^{-\frac{1}{2}}\, S^{-\frac{1}{2}}\, u = \lambda u$

Prémultiplions les deux membres par $S^{-\frac{1}{2}}$, et posons $S^{-\frac{1}{2}}\, u = v$.

Il vient : $$S^{-\frac{1}{2}}\, A\, S^{-\frac{1}{2}}\, v = \lambda v$$

La matrice $Q'' = S^{-\frac{1}{2}}\, A\, S^{-\frac{1}{2}}$ est symétrique, et a même valeur propre λ que Q.

Il suffit donc de la diagonaliser, puis de calculer u par la

relation $u = S^{\frac{1}{2}} v$.

Le facteur $\Psi = S^{-1} u$ vaut : $\varphi = S^{-\frac{1}{2}} v$.

c) Pratique des calculs

On suppose $p \leq n$.

1 - On calcule la matrice simplifiée Q" : $q''_{jj'} = \sum_{i=1}^{n} \frac{1}{f_{i.}} \frac{f_{ij} f_{ij'}}{\sqrt{f_{.j} f_{.j'}}}$

(Q" est d'ordre p, p)

2 - On calcule ses vecteurs et valeurs propres (en négligeant le vecteur propre relatif à $\lambda = 1$).

3 - Les facteurs s'obtiennent par la transformation $S^{-\frac{1}{2}} = \left(\frac{\delta_{ij}}{\sqrt{f_{.j}}} \right)$

à partir des vecteurs propres. Ils fournissent les coordonnées des p points de l'autre nuage.

4 - On calcule les projections des n points du nuage par la formule :

$$\Psi_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} \varphi_j$$

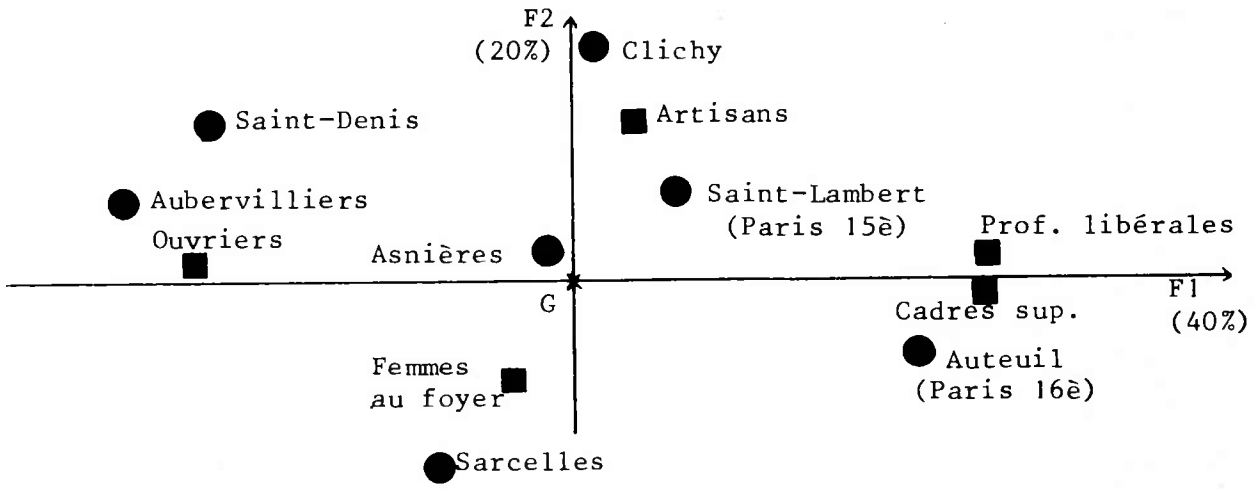
5 - On réalise le graphique du plan des deux premiers facteurs correspondant à $\lambda \neq 1$.

II-3. INTERPRETATION DES RESULTATS

II-3.1 - Généralités

Comme dans le cas de l'analyse en composantes principales, l'essentiel des résultats sera condensé dans les graphiques représentant les nuages dans les plans formés par les premiers axes factoriels pris deux à deux.

Figure 5



La figure 5 nous donne l'exemple de 5 catégories d'activité, représentées sur le même graphique que 7 communes ou quartiers de la Région Parisienne.

Nous allons utiliser cet extrait extrêmement simplifié de l'analyse globale qui nous a servi d'exemple le long de l'exposé pour mettre en évidence les règles d'interprétation de l'analyse des correspondances.

Les pourcentages qui figurent sur les axes représentent la part d'inertie (ou de variance) "expliquée" par ces axes : ainsi F_1(40%) signifie que la première valeur propre représente 40% de la somme de toutes les valeurs propres. Les deux premiers axes ci-dessus expliquent donc ensemble 40 + 20 = 60% de l'inertie totale. Ce pourcentage donne une idée (pessimiste) de la part d'information fournie par les facteurs. Nous disons "pessimiste" car cette façon de mesurer l'information est extrêmement partielle ; il arrive que les pourcentages soient faibles et que, néanmoins, les facteurs correspondants restituent l'essentiel de l'information initiale.

Comme dans le cas de l'A.C.P., il sera licite d'interpréter les proximités entre éléments d'un même nuage ; ainsi, Aubervilliers et Saint-Denis ont des profils d'activités voisins, les professions libérales et les cadres supérieurs ont des profils géographiques

voisins ; il sera également licite d'interpréter les positions relatives de deux points d'un ensemble par rapport à tous ceux de l'autre ensemble. Sauf cas particulier, il est extrêmement périlleux d'interpréter la proximité de deux points correspondant à des nuages différents.

Le centre de gravité G, qui est à l'origine des axes, correspond aux profils moyens (cf. § II-3). Ainsi, le profil socio-professionnel d'Asnières est voisin du profil socio-professionnel moyen de toute la région parisienne.

II-3.2 - Calcul des contributions absolues et relatives.

Pour l'interprétation des axes, il est utile de calculer deux séries de coefficients, pour chaque facteur :

a) les contributions absolues qui exhibent la part prise par une variable dans l'inertie (ou variance) expliquée par un facteur : cette part va être calculée par rapport à l'ensemble des variables.

b) les contributions relatives qui exhibent la part de la dispersion d'une variable expliquée par un facteur.

[N.B. - Ces deux notions distinctes coïncident, à de légères modifications près, dans le cas de l'analyse en composantes principales avec celle de coefficient de corrélation variable-composante.

En effet, l'inertie expliquée par le facteur q vaut λ_q, où λ_q désigne maintenant la q-ème valeur propre de la matrice des corrélations.

Or les coordonnées du point-variable "i" sur les nouveaux axes sont proportionnelles à $u_{iq}\sqrt{\lambda_q}$ et l'on a bien : $\sum_i (u_{iq}\sqrt{\lambda_q})^2 = \lambda_q$, u étant unitaire. Le nombre $100\ u_{iq}^2$ exprime bien le pourcentage de variance expliquée par la variable i.

D'autre part, la distance d'une variable "i" à l'origine dans R^n vaut 1 (cf. § IV-2 ACP). On a donc, dans la base orthonormée des axes factoriels : $\sum_q \lambda_q u_{iq}^2 = 1$.

La quantité $\lambda_q u_{iq}^2$ exprime bien la part de l'axe q dans la variance de la variable i].

a) Contributions absolues

La norme du facteur φ_q vaut 1 :

$$\sum_{j=1}^{p} f_{.j} \varphi_{jq}^2 = 1$$

La projection du point "j" de R^n sur l'axe factoriel "q" vaut:

$$\sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} \psi_{iq} = \sqrt{\lambda_q}\, \varphi_{jq}$$

La variance du nuage projeté vaut donc :

$$\sum_{j} (\sqrt{\lambda_q}\, \varphi_{jq})^2 \cdot f_{.j} = \lambda_q$$

Le quotient $\dfrac{\lambda_q \varphi_{jq}^2 f_{.j}}{\lambda_q} = f_{.j} \varphi_{jq}^2 = ca_q(j)$ représente la contribution de la variable j au facteur q.

Notons que $\sum_{j} ca_q(j) = 1$

b) Contributions relatives

Dans R^n, le carré de la distance au centre de gravité de la variable j vaut :

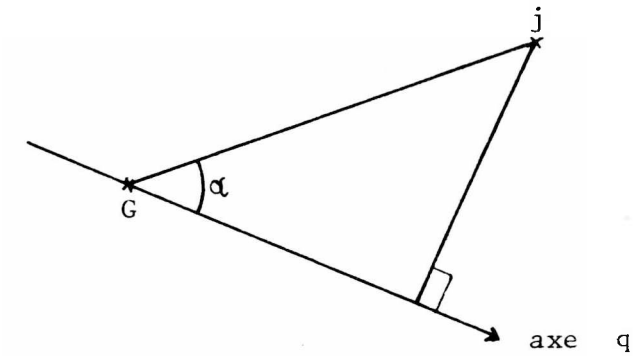
$$d^2(j, G) = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2$$

Le carré de la projection de la variable j sur l'axe q vaut :

$$d_q^2 (j, G) = (\sqrt{\lambda_q}\ \varphi_{jq})^2$$

Notons que $\sum_q d_q^2(j, G) = d^2(j, G)$

Figure 6



La quantité : $\cos^2 \alpha = \dfrac{d_q^2(j, G)}{d^2(j, G)} = cr_q(j)$ représente bien la part qui revient au facteur q dans l'explication de la variance de la variable j.

Notons que $\sum_q cr_q(j) = 1$

Tout ce qui vient d'être dit sur les p variables dont les projections sur le q-ème axe valent $[(\sqrt{\lambda_q}\ \varphi_{jq}),\ j = 1, \ldots, p]$ peut être transposé de façon symétrique pour les n variables de l'autre ensemble, dont les projections sur le q-ème axe factoriel valent : $[(\sqrt{\lambda_q}\ \psi_{iq}),\ i = 1, \ldots, n]$.

c) Exemple d'utilisation

(Les coefficients ca_q et cr_q sont multipliés par 100, afin d'être interprétés en termes de pourcentages).

Supposons que le point j : "cadre supérieur" ait une contribution absolue de 5% et une contribution relative de 75% au premier axe.

Le point "cadre supérieur" n'est donc pas intervenu fortement lors de la "construction" de cet axe (comme ce point est éloigné sur l'axe, cette faiblesse de la contribution ne peut être due qu'à la faiblesse de la masse $f_{.j}$ de la variable j).

Par contre, le premier axe explique 75% de la variable "j : cadre supérieur" (ou encore : $\cos^2\alpha = 0,75$, ce qui signifie que la variable fait un angle très faible avec l'axe).

Ainsi, la variable "cadre supérieur" est une caractéristique exclusive du premier axe.

Supposons que le point i : "Aubervilliers" ait une contribution absolue de 30% et une contribution relative de 10%. Cela signifie qu'il participe fortement à la construction du premier axe, mais qu'il participe probablement à celle de nombreux autres axes. (La masse $f_{i.}$ et la distance à l'origine $d^2(i, G)$ sont importantes).

II-4. AUTRE PRESENTATION DE L'ANALYSE DES CORRESPONDANCES : RECHERCHE DE LA MEILLEURE REPRESENTATION SIMULTANEE

Nous reprendrons l'exemple du tableau de contingence X d'ordre (n,p), croisant les n communes de la région parisienne avec p catégories d'activité (notées CSP).

Nous allons chercher à représenter sur un même axe l'ensemble des communes et l'ensemble des CSP, de façon à approcher au mieux la situation idéale suivante :

A chaque point CSP "j" est un barycentre des points-communes, chaque commune "i" étant affectée du poids : "Part de la commune i dans la CSP j", autrement dit, du poids : $p_i = f_{ij}/f_{.j}$ $(\sum_j p_i = 1)$

B chaque point-commune "i" est un barycentre des points-CSP, chaque CSP "j" étant affectée du poids : "Part de la CSP j dans la commune i", autrement dit, du poids : $p'_j = f_{ij}/f_{i.}$ $(\sum_j p'_j = 1)$.

Cette situation idéale est en général impossible, car elle implique que

chaque ensemble est contenu dans l'autre. (Il existe une solution triviale, pour laquelle tous les points des deux ensembles sont confondus avec le point d'abscisse 1).

Avec les notations du § I.4, si φ_j désigne l'abscisse de la CSP j sur l'axe, (φ_j étant la j-ème composante d'un vecteur φ), si ψ_i désigne l'abscisse de la commune i sur ce même axe, les conditions A et B s'écrivent respectivement :

$$\text{A} \qquad \varphi = S^{-1} \tilde{F} \psi$$

$$\text{B} \qquad \psi = T^{-1} F \varphi$$

Ces relations strictement barycentriques étant en général impossibles à réaliser simultanément, nous chercherons un coefficient $\alpha > 0$ le plus proche de 1, tel que l'on ait les relations :

$$\text{A'} \qquad \varphi = \alpha S^{-1} \tilde{F} \psi$$

$$\text{B'} \qquad \psi = \alpha T^{-1} F \varphi$$

Remarquons que α est forcément plus grand ou égal à 1, sinon les relations A' et B' impliqueraient encore que chacun des deux ensembles recouvre un intervalle de l'axe strictement contenu dans l'intervalle recouvert par l'autre ensemble.....

On est donc conduit à chercher le plus petit α tel que A' et B' soient vérifiées.

Remplaçant par exemple, dans B', φ par sa valeur tirée de A' :

$$\psi = \alpha^2 T^{-1} F S^{-1} \tilde{F} \psi$$

autrement dit, ψ est vecteur propre de $T^{-1} F S^{-1} \tilde{F}$ relatif à la plus grande valeur propre : $\lambda = 1/\alpha^2$.

Les relations A' et B', où $\alpha = 1/\sqrt{\lambda}$, ne sont autres que les relations (3) et (4) du § II.2.

φ et ψ sont les premiers facteurs correspondant aux deux ensembles mis en correspondance.

On peut étendre cette recherche de la meilleure représentation "α-barycentrique" sur un axe, à celle de la meilleure représentation "α, β barycentrique" dans un plan repéré par deux axes orthogonaux, puis généraliser à un sous-espace de dimension quelconque (inférieure à p). On retrouve bien entendu la représentation déjà obtenue par l'analyse des correspondances.

II-5. PRESENTATION COMME CAS PARTICULIER DE L'ANALYSE DISCRIMINANTE

L'analyse discriminante, brièvement exposée au paragraphe V, tente d'exhiber une partition "a priori" des individus à partir de variables mesurées sur ces individus - Si ces variables sont des variables indicatrices d'une partition, l'analyse discriminante coïncide alors avec l'analyse des correspondances du tableau de contingence croisant les deux partitions.

On appelle ici variables indicatrices d'une partition en p classes d'un ensemble E de cardinal m un ensemble de p variables $Z_1, \ldots\ldots Z_p$ telles que $Z_{ij} = 1$ si l'individu i appartient à la classe j, d'effectif $m_{j.}$, $Z_{ij} = 0$ sinon. Les propriétés des partitions impliquent alors que l'on ait les relations :

$$\sum_{i=1}^{m} Z_{ij} Z_{ij'} = \delta_{jj'} m_{j.} \quad ; \quad \sum_{i=1}^{m} \sum_{j=1}^{p} Z_{ij} = m$$

En fait, les deux partitions joueront des rôles symétriques. L'analyse discriminante est un cas particulier de l'analyse canonique lorsque l'un des deux ensembles de variables est constitué des variables indicatrices d'une partition : nous avons donc ici affaire en quelque sorte à une "double analyse discriminante".

Soit Z le tableau de données à m lignes (individus) et p colonnes. Z_{ij} vaut 1 si l'individu i appartient à la classe j, 0 sinon. Une ligne de Z ne contient donc que des "0", et un seul "1".

Le tableau analysé a donc p colonnes et m lignes, réparties par ailleurs en q classes.

En adoptant des notations analogues à celles du paragraphe V nous allons calculer les moyennes générales, les moyennes de classes, les matrices des covariances interclasses et intraclasses relatives aux variables indicatrices.

Nous noterons f_{jk} la fréquence relative et m_{jk} la fréquence absolue des individus appartenant à la cause de décès j et au département k, avec :

$$f_{.k} = \sum_j f_{jk} = \frac{m_{.k}}{m} \qquad \text{(fréquence du département } k\text{)}$$

$$f_{j.} = \sum_k f_{jk} = \frac{m_{j.}}{m} \qquad \text{(fréquence de la cause de décès } j\text{)}$$

$$f_{jk} = \frac{m_{jk}}{m}$$

La moyenne générale de la variable j s'écrit :

$$\overline{Z}_j = \frac{1}{m} \sum_{i=1}^{m} Z_{ij} = \frac{m_{j.}}{m} = f_{j.}$$

La moyenne de la variable j pour la classe k caractérisée par un ensemble I_k d'indices, vaut :

$$\overline{Z}_{kj} = \frac{1}{m_{.k}} \sum_{i \in I_k} Z_{ij} = \frac{m_{jk}}{m_{.k}} = \frac{f_{jk}}{f_{.k}}$$

La matrice des covariances globale T a pour terme général :

$$t_{jj'} = \frac{1}{m} \sum_{i=1}^{m} \left(Z_{ij} - \overline{Z}_j\right) \left(Z_{ij'} - \overline{Z}_{j'}\right) = \frac{1}{m} \sum_i Z_{ij} Z_{ij'} - \overline{Z}_j \overline{Z}_{j'}$$

Soit $$t_{jj'} = \delta_{jj'} f_{j.} - f_{j.} f_{j'.}$$

La matrice des covariances interclasse E a pour terme général :

$$e_{jj'} = \sum_{k=1}^{q} \frac{m_{.k}}{m} \left(\overline{Z}_{jk} - \overline{Z}_j\right) \left(\overline{Z}_{j'k} - \overline{Z}_{j'}\right)$$

$$= \sum_{k=1}^{q} f_{.k} \left(\frac{f_{jk}}{f_{.k}} - f_{j.}\right) \left(\frac{f_{j'k}}{f_{.k}} - f_{j'.}\right)$$

La première fonction discriminante u vérifie la relation, pour λ maximal :

$$Eu = \lambda Tu$$

Une petite difficulté de calcul vient de ce que la matrice T est singulière. (La somme des éléments de ses lignes ou de ses colonnes est nulle).

Appelons S la matrice de terme général $\delta_{jj'} f_{j.}$, et f le vecteur dont la j-ème composante vaut $f_{j.}$. On a alors :

$$T = S - f\tilde{f}$$

Résolvons tout d'abord l'équation, pour λ maximal :

$$E\varphi = \lambda S\varphi$$

La matrice $S^{-1}E$ n'est autre que la matrice Q' définie plus haut (§II-2.3). Les vecteurs propres φ ne sont autres que les facteurs de l'analyse des correspondances du tableau de terme général f_{ik}.

Pour un facteur φ correspondant à une valeur propre non nulle de Q', on a en particulier la relation $\tilde{f}\varphi = 0$, qui exprime que ce facteur est centré :

$$\text{Soit} \qquad \sum_j \varphi_j f_{j.} = 0$$

Il s'ensuit que les $p-1$ vecteurs propres de $S^{-1}E$, qui vérifient la relation :

$$E\varphi = \lambda S\varphi \qquad \text{avec} \quad \tilde{p}'\varphi = 0$$

vérifient également : $E\varphi = \lambda(S - p\tilde{p})\varphi = \lambda T\varphi$

Ils coïncident donc avec les $p-1$ fonctions discriminantes.

En intervertissant les rôles joués par les deux partitions, on constate de la même façon que les $q-1$ nouvelles fonctions discriminantes coïncident avec les $q-1$ facteurs relatifs à l'autre ensemble.

Cette présentation de l'analyse des correspondances, qui est surtout un exercice, permet néanmoins de donner une interprétation intéressante de la

plus grande valeur propre, comme "pouvoir discriminant" des facteurs vis-à-vis des partitions étudiées.

III - ANALYSE FACTORIELLE CLASSIQUE

III - 1. HISTORIQUE. EVOLUTION DU MODELE DE BASE

Nous désignons par "Analyse factorielle classique" la technique mise au point essentiellement par SPEARMAN (1904), puis développée par THURSTONE et rattachée à la statistique mathématique par LAWLEY et MAXWELL.

Cette technique, dont les applications concernent essentiellement la psychologie, diffère essentiellement de l'analyse en composantes principales (dont elle est à l'origine) parce qu'elle suppose l'existence d'un modèle "a priori" qui peut schématiquement se résumer de la façon suivante:

"Les différentes variables mesurées sur un même individu dépendent d'un très petit nombre de facteurs (ou variables latentes, cachées) indépendants, communs à toutes les variables, et de facteurs spécifiques à chaque variable, ceux-ci étant indépendants des facteurs communs, et indépendants entre eux".

Avant de préciser cette définition d'ailleurs vague et imprécise du modèle, il peut être utile de retracer brièvement son évolution historique.

Bien qu'entrevu par Karl PEARSON, le premier modèle d'analyse factorielle fut posé par le psychologue SPEARMAN : les différentes notes à des tests psychologiques ne faisaient que refléter, selon lui, une certaine aptitude générale du sujet interrogé, dont la mesure directe est impossible.

C'est le modèle de l'analyse unifactorielle, que l'on formalise de la manière suivante : $\vec{x}_i$ désignant le vecteur des p variables mesuré sur l'individu i (dont les composantes sont notées : $x_{1i}, x_{2i}, \ldots, x_{pi}$) $\vec{a}$ désignant un vecteur de coefficients à p composantes, et e_i un vecteur résiduel :

$$\vec{x}_i = \vec{a} f_i + \vec{e}_i \qquad i = 1, 2, \ldots, p$$

Ainsi, le vecteur des p notes $\vec{x}_i$, pour l'individu i, ne dépend que de la note scalaire f_i (qui représente par exemple une note d'intelligence), à un résidu aléatoire près $\vec{e}_i$. Le vecteur $\vec{a}$ est le même pour tous les individus.

Les composantes de $\vec{a}$ sont appelées les saturations des différentes variables. Elles caractérisent la façon dont les variables effectivement mesurées dépendent du facteur général.

En nous référant à une population théorique, le modèle s'exprime de la façon suivante : le vecteur aléatoire $\vec{x}$ est fonction de la variable aléatoire f et du vecteur aléatoire $\vec{e}$ selon la formule : $\vec{x} = \vec{a}\, f + \vec{e}$.

Les composantes de $\vec{e}$ sont indépendantes entre elles et indépendantes de f.

Le modèle s'est ensuite affiné en "analyse bifactorielle", où deux facteurs sous-jacents, f_1 et f_2, indépendants, suffisaient à reconstituer les notes, à un résidu aléatoire près : $\vec{x} = \vec{a}_1 f_1 + \vec{a}_2 f_2 + \vec{e}$.

Enfin, ce fut la formulation générale de l'analyse multifactorielle, où le nombre de variables sous-jacentes n'est plus fixé à l'avance, mais suggéré (de façon souvent empirique et approximative) par les calculs eux-mêmes.

$$\vec{x} = \vec{a}_1 f_1 \ldots\ldots + \vec{a}_r f_r + \vec{e}$$

ou encore :

$$x = A\, f + e$$

x = matrice colonne des composantes de x

A = matrices dont la i-ème colonne est formée des composantes de $\vec{a}_i$

e = matrice colonne des composantes de $\vec{e}$

a) Les composantes de $\vec{e}$ sont supposées non corrélées et de moyennes nulles, autrement dit, E désignant le symbole "Espérance mathématique" :

$E(e\, \tilde{e}) = \triangle$ ($\triangle$ = matrice diagonale d'ordre p, p)

(puisque $E(e_i e_j) = 0$ si $i \neq j$)

b) Les différents "facteurs" $f_1, f_2, \ldots\ldots, f_n$, composantes de f, sont généralement supposés centrés, de corrélation nulle et de va-

riance 1 :

$$E(f\,\tilde{f}) = I_r \qquad (I_r = \text{matrice unité d'ordre } r)$$

c) Les facteurs et les termes résiduels sont supposés indépendants :

$$E(f\,\tilde{e}) = 0 \qquad (0 = \text{matrice nulle d'ordre } r, p)$$

d) Les composantes du vecteur x (qui est le seul vecteur aléatoire dont on observera effectivement des réalisations) sont supposées de moyenne nulle.

III-2. PRINCIPES DES CALCULS

Calculons la matrice des covariances du vecteur x :

$$V(x) = E(x\,\tilde{x}) = E\left[(A\,f + e)\widetilde{(A\,f + e)}\right]$$
$$= E\left[(Af + e)(\tilde{f}\,\tilde{A} + \tilde{e})\right]$$

Soit :

$$V(x) = E\left[A\,f\,\tilde{f}\,\tilde{A} + e\,\tilde{f}\,\tilde{A} + A\,f\,\tilde{e} + e\,\tilde{e}\right]$$

Compte tenu de la linéarité de l'opérateur E :

$$= A\,E(f\,\tilde{f})\,\tilde{A} + E(e\,\tilde{f})\,\tilde{A} + A\,E(f\,\tilde{e}) + E(e\,\tilde{e})$$

Or : $E(f\,\tilde{f}) = I_r$, $E(f\,\tilde{e}) = 0$, d'où $E(e\,\tilde{f}) = 0$, et $E(e\,\tilde{e}) = \Delta$

On en déduit la relation fondamentale de l'analyse factorielle classique qui s'écrit , en indiquant les ordres des matrices sous chacune d'elles :

$$\underset{(p,\,p)}{E(x\,\tilde{x})} = \underset{(p,r)(r,p)}{A\,\tilde{A}} + \underset{(p,p)}{\Delta} \qquad (1)$$

En pratique, on ne dispose que de n réalisations du vecteur x, et donc d'une estimation de la matrice des covariances $E(x\,\tilde{x})$ que l'on notera V^*.

Par contre, tout le membre de droite de la relation (1) est inconnu.

D'autre part, on désire généralement que le nombre r de facteurs soit le plus faible possible ; on désire également que les variances résiduelles, éléments diagonaux de Δ, ne soient pas trop importantes.

Le problème de calcul statistique sera donc le suivant, en gardant une

formulation intuitive :

- une estimation V^* de la matrice des covariances du vecteur aléatoire x étant donnée, peut-on trouver une matrice A^* , ayant le moins de colonnes possible, et une matrice diagonale à éléments positifs Δ^* telle que l'on ait la relation :

$$V^* \simeq A^* \tilde{A}^* + \Delta^* \qquad (2)$$

Il est inutile, pensons-nous, d'insister sur les difficultés que pose la formulation précise et les divers essais de résolution de ce problème statistique ; il est cependant aisé de percevoir l'origine et l'étendue de ces difficultés.

a) Nous avons à estimer simultanément $r \times p + p = (r + 1)\, p$ paramètres, uniquement à partir de relations implicites entre ces paramètres.

b) Les fluctuations d'échantillonnages peuvent dégrader considérablement la qualité de la relation théorique : $V = A \tilde{A} + \Delta$.

c) Le modèle est en fait indéterminé, car, si T désigne une matrice p, p orthogonale, et si A^* est une matrice satisfaisant la relation (2) avec Δ^* , alors A^*T et Δ^* vérifient également la relation (2) (puisque $T \tilde{T} = I_p$).

III-3. CAS PARTICULIERS ET RESOLUTION DU PROBLEME

III-3.1 - Variances spécifiques nulles.

Supposons que la matrice diagonale Δ ait tous ses éléments égaux à 0. Ceci implique que le vecteur résiduel e soit constant, et donc nul, puisque ses composantes sont de moyennes nulles.

On doit donc chercher une matrice A^* telle que :

$$V^* = A^* \tilde{A}^*$$

Ceci n'est autre que la décomposition réalisée en analyse en composantes principales. En effet, soit U la matrice d'ordre (p, p) dont les colonnes sont les p vecteurs propres de V^* ,

classés selon des valeurs propres décroissantes. Soit A la matrice diagonale des valeurs propres de V^* classées de la même façon.

On a la relation : $V^* = U \Lambda \tilde{U}$

Posons : $\Lambda^{\frac{1}{2}} = (\sqrt{\lambda_{ij}})$ (ou $\lambda_{ij} = 0$ si $i \neq j$)

On a alors :

$$V^* = U \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} \tilde{U} = U \Lambda^{\frac{1}{2}} (\widetilde{U \Lambda^{\frac{1}{2}}}) = A_1^* \tilde{A}_1^*$$

(où l'on a posé : $A_1^* = U \Lambda^{\frac{1}{2}}$)

Remarquons que si les $p-r$ dernières valeurs propres de V^* sont très voisines de 0, les $p-r$ dernières colonnes de $A_1 = U \Lambda^{\frac{1}{2}}$ sont quasi-nulles, ce qui nous conduit à la solution, en désignant par A^* la matrice formée des r premières colonnes de A_1 :

$$\underset{(p,p)}{V^*} \simeq \underset{(p,r)}{A^*} \underset{(r,p)}{\tilde{A}^*} \qquad (3)$$

Ainsi, si les $p-r$ dernières valeurs propres de V sont considérées comme négligeables, les r premières composantes principales constituent r facteurs indépendants, permettant de reconstituer non seulement la matrice V^* par la relation (3) mais également les observations de départ x_i, $(i = 1, \ldots, n)$ par la formule :

$$x_i = A^* f_i$$

Nous verrons que lorsqu'il existe des facteurs spécifiques e_i non nuls, les facteurs permettent de reconstituer seulement la matrice V^*, et par conséquent les corrélations entre variables.

III-3.2 - Variances spécifiques égales.

Supposons que la matrice Δ soit de la forme : $\Delta = \sigma^2 I$.

Ainsi, dans le modèle $x = A f + e$, chaque composante de e a une variance σ^2. La relation fondamentale (1) s'écrit :

$$V = A \tilde{A} + \sigma^2 I \qquad (4)$$

Supposons que σ^2 soit connu. L'équivalent empirique de la relation (4) s'écrit :

$$V^* = A^* \tilde{A}^* + \sigma^2 I$$

On peut donc, en cherchant les composantes principales de $V^* - \sigma^2 I$, qui est connue, nous ramener au cas des variances spécifiques nulles.

Nous avions, au paragraphe précédent : $V^* = U \Lambda \tilde{U}$

Soit, puisque $U \tilde{U} = I$

$$V^* - \sigma^2 I = U \Lambda \tilde{U} - \sigma^2 U \tilde{U} = U(\Lambda - \sigma^2 I) \tilde{U}$$

La matrice $\Lambda - \sigma^2 I$ est diagonale ; son i-ème terme diagonal vaut $\lambda_i - \sigma^2$. Autrement dit, $V^* - \sigma^2 I$ a mêmes vecteurs propres que V^*, et des valeurs propres obtenues en retranchant σ^2 à celles de V^*.

Ceci nous fournit à la fois une méthode de calcul des facteurs et d'estimation de σ^2.

En effet, si les $p-r$ plus petites valeurs propres de V^* peuvent être considérées comme <u>égales</u> (à des fluctuations d'échantillonnage près)à une même valeur s^2,alors les $p-r$ plus petites valeurs propres de $V^* - s^2 I$ peuvent être considérées comme <u>nulles</u>, et par conséquent, il existe une matrice A^* d'ordre (p, r) telle que :

$$V^* - s^2 I = A^* \tilde{A}^*$$

Ainsi, pour que les données puissent être considérées comme générées par un modèle à r facteurs communs et à variances spécifiques égales, il suffit que les $p-r$ plus petites valeurs propres de V^* soient statistiquement égales

On estimera la variance spécifique s^2 par la moyenne arithmétique de ces $p-r$ valeurs propres.

Il existe des tests statistiques destinés à éprouver l'éga-

lité des dernières valeurs propres de V^*; ils sont en fait assez peu usités, car ils font appel à des hypothèses de départ assez peu souvent réalisées dans la pratique.

III-3.3 - Cas général : solutions approchées.

Il existe de nombreuses techniques d'estimation des paramètres du modèle. Beaucoup d'entre elles, comme la "méthode centroïde de THURSTONE" bien connue des psychologues, étaient particulièrement bien adaptées aux calculs manuels.

D'autres, comme la méthode du maximum de vraisemblance, proposées par LAWLEY et MAXWELL, ont des fondements statistiques plus solides, mais ne convergent pas toujours de façon satisfaisante.

Nous donnons ci-dessous une méthode qui donne des résultats satisfaisants, et dont le principal mérite est d'être simple : l'Analyse en facteurs principaux .

Soit donc à trouver A^* et Δ^* tels que :

$$V^* = A^* \tilde{A}^* + \Delta^*$$

On commence par estimer Δ^*.

On peut par exemple poser $\Delta^* = 0$. Cependant, la convergence a lieu plus rapidement si l'on pose :

$$\Delta^* = \left[\text{diag}(V^{*-1}) \right]^{-1}$$

où $\text{diag}(V^{*-1})$ désigne la matrice diagonale ayant les mêmes éléments diagonaux que V^{*-1}. Cela revient à assimiler la variance spécifique relative à une variable à la variance résiduelle de la régression multiple expliquant cette variable par toutes les autres.

Ayant estimé Δ^* par D_o^*, on cherche les composantes principales de $V^* - D_o^*$, et l'on obtient, comme dans le cas des variances spécifiques nulles, une décomposition de la forme

$$V^* - D_o^* \simeq B \tilde{B} \qquad B \text{ étant d'ordre } (p, r)$$

(après avoir négligé les $p-r$ plus petites valeurs propres).

On estime de nouveau Δ^* par $D_1^* = \text{diag}(V^* - B\,\tilde{B})$ (éléments diagonaux de $V^* - B\,\tilde{B}$).

Puis on décompose de nouveau $V^* - D_1^*$, ce qui nous donne une nouvelle estimation D_2^*, etc... On réitère jusqu'à l'obtention d'une certaine stabilité de Δ^*.

En l'absence de critères solides, il est fréquent de se fonder sur des considérations empiriques pour choisir le nombre exact de facteurs et la qualité de l'approximation.

III-3.4 - Rotations et axes obliques.

Nous avons vu que les solutions étaient indéterminées, et que toute transformation orthogonale (ou rotation) sur les facteurs fournit une nouvelle solution.

D'où l'idée des psychologues de procéder à des rotations de façon à obtenir des facteurs dont l'interprétation soit aisée.

Différents critères existent et sont en concurrence pour "améliorer" les résultats bruts tels qu'ils sont définis par un algorithme tel que celui du paragraphe précédent.

De plus, nous avons jusqu'ici fait l'hypothèse que les facteurs étaient indépendants (en fait, de corrélations nulles). Cette hypothèse avait l'avantage d'être simple et de lever une partie de l'indétermination du modèle.

Cependant, l'hypothèse de non-corrélation rigoureuse des facteurs est assez irréaliste si l'on se propose d'interpréter les facteurs à partir de concepts existants (facteur-intelligence et facteur-mémoire, par exemple).

Nous serions ainsi conduits à explorer toute une famille de techniques qui constituent la partie la moins spécifique et la plus contestée de l'analyse factorielle classique, bien qu'elles donnent d'excellents résultats lorsqu'elles sont manipulées par des mains expertes.

Pour ces développements, nous renvoyons le lecteur à l'ouvrage "Modèles et Méthodes de l'Analyse Factorielle" de J. TORRENS-IBERN (Dunod 1972), consacré exclusivement à l'analyse factorielle classique, qui en constitue pratiquement le seul exposé synthétique en langue française.

IV - ANALYSE CANONIQUE

La méthode d'analyse canonique, développée par HOTELLING (ou encore analyse des corrélations canoniques) présente un intérêt assez limité pour les applications, car elle conduit à de grandes difficultés d'interprétation. Cependant, elle joue un rôle théorique important : en effet, elle constitue un cadre général dont la régression multiple, la plupart des techniques d'analyse des données exposées précédemment, et l'analyse discriminante exposée plus loin sont des cas particuliers. L'analyse canonique cherche à synthétiser les interrelations existant entre deux groupes de variables, en cherchant les combinaisons linéaires des variables du premier groupe les plus corrélées à des combinaisons linéaires des variables du second groupe. On retrouve immédiatement la régression multiple si l'un des deux groupes n'est constitué que par une seule variable.

IV-1. NOTATIONS ET FORMULATION DU PROBLEME

Le tableau de données D, à n lignes et $p+q$ colonnes, est partitionné en deux sous-tableaux X et Z, ayant respectivement p et q colonnes

$$D = (X,Z)$$

Les lignes représentent les individus ou observations, les p premières colonnes sont les variables du premier groupe, et les q suivantes sont celles du second groupe.

Nous supposerons que les variables sont centrées, ce qui signifie que chaque colonne D est telle que la somme de ses éléments vaut 0. Ainsi, la matrice des covariances expérimentales des $p+q$ variables s'écrit :

$$V(D) = \frac{1}{n} \tilde{D}\, D \qquad \left(v_{ij} = \frac{1}{n} \sum_k d_{ki} d_{kj} \right)$$

$$V(D) = \frac{1}{n} \begin{bmatrix} \tilde{X}\, X & \tilde{X}\, Z \\ \tilde{Z}\, X & \tilde{Z}\, Z \end{bmatrix}$$

Ce qui peut encore s'écrire, en faisant intervenir les matrices des covariances internes à chaque groupe de variables : $V_{XX} = 1/n\ \tilde{X}\ X$ et $V_{ZZ} = 1/n\ \tilde{Z}\ Z$ ainsi que le tableau des covariances entre les groupes :

$$V_{XZ} = \frac{1}{n}\ \tilde{X}\ Z \quad (\text{avec} \quad V_{ZX} = \tilde{V}_{XZ})$$

$$V(D) = \begin{bmatrix} V_{XX} & V_{XZ} \\ V_{ZX} & V_{ZZ} \end{bmatrix}$$

Intéressons-nous à l'individu k, caractérisé par les variables : (k-ème ligne de D)

$$x_{k1},\ x_{k2},\ \ldots\ x_{kp},\ z_{k1},\ z_{k2},\ \ldots\ z_{kq}$$

Soient a et b deux vecteurs à p et q composantes, définissant deux combinaisons linéaires a(k) et b(k) :

$$a(k) = \sum_{i=1}^{p} a_i x_{ki}$$

$$b(k) = \sum_{j=1}^{q} b_j z_{kj}$$

Les n valeurs de a(k) pour tous les individus sont les n lignes de Xa. De même, les n valeurs de b(k) sont les lignes de Zb.

Puisque les variables initiales sont centrées, leurs combinaisons linéaires sont également centrées.

Nous nous proposons de chercher les deux combinaisons linéaires a(k) et b(k) les plus corrélées sur l'ensemble des valeurs de k. Elles prendront alors le nom de variables canoniques.

Comme le coefficient de corrélation ne dépend pas de l'échelle des variables, nous imposerons aux deux combinaisons linéaires d'avoir variance 1.

Calculons tout d'abord la variance de a(k), notée brièvement var(a)

$$\text{var}(a) = \frac{1}{n} \sum_{k=1}^{n} a^2(k) = \frac{1}{n} (\tilde{Xa})\ Xa = \frac{1}{n} \tilde{a}\ \tilde{X}\ Xa$$

Soit : $\text{var}(a) = \tilde{a}\ V_{XX} a = 1$

De la même façon :

$$\text{var}(b) = \tilde{b}\ V_{ZZ} b = 1$$

Dans ces conditions, le coefficient de corrélation entre les combinaisons linéaires $a(k)$ et $b(k)$ s'identifie avec la covariance :

$$\text{cov}(a,b) = \frac{1}{n} \sum_{k=1}^{n} a(k).b(k)$$

Soit :

$$\text{cov}(a,b) = \frac{1}{n} \tilde{a}\ \tilde{X}\ Zb = \tilde{a}\ V_{XZ} b$$

Le problème mathématique est le suivant :

Rendre maximal $\text{cov}(a,b) = \tilde{a}\ V_{XZ} b$
avec les contraintes $\tilde{a}\ V_{XX} a = \tilde{b}\ V_{ZZ} b = 1$

IV-2. CALCUL DES VARIABLES CANONIQUES

La démonstration est analogue à celle de l'analyse générale. Deux multiplicateurs de LAGRANGE interviennent.

Il nous faut rendre maximal : $\mathcal{L} = \tilde{a}\ V_{XZ} b - \lambda(\tilde{a}\ V_{XX} a - 1) - \mu\ (\tilde{b}\ V_{ZZ} b - 1)$
D'où le système :

$$V_{XZ} b - 2\lambda V_{XX} a = 0 \qquad (A)$$

$$V_{ZX} a - 2\mu V_{ZZ} b = 0 \qquad (B)$$

Prémultiplions les deux membres des relations (A) et (B) respectivement pas $\tilde{a}$ et $\tilde{b}$, puis tenant compte des contraintes :

$$\tilde{a}\ V_{XX} a = b' V_{ZZ} b = 1$$

Il vient : $\tilde{a}\, V_{XZ} b = 2\lambda$

$\tilde{b}\, V_{ZX} a = 2\mu$

D'où, puisque les deux premiers membres sont des scalaires transposés, donc égaux : $\lambda = \mu$.

Notons que la valeur commune des multiplicateurs de LAGRANGE notée désormais λ est deux fois la valeur du maximum cherché.

La relation (A) nous donne a (si V_{XX} est inversible, pour $\lambda \neq 0$)

$$a = \frac{1}{2\lambda} V_{XX}^{-1} V_{XZ}\, b$$

En reportant la valeur de a dans la relation (B) et en posant $\beta = 2\lambda$

$$V_{ZX} V_{XX}^{-1} V_{XZ} b = \beta^2 V_{ZZ} b \qquad (C)$$

Ce qui prouve que b est vecteur propre de $V_{ZZ}^{-1} V_{ZX} V_{XX}^{-1} V_{XZ}$ relatif à la plus grande valeur propre β^2, qui est le carré du coefficient de corrélation entre les combinaisons linéaires a(k) et b(k).

β^2 est appelé première racine canonique, ou premier coefficient de corrélation canonique entre les deux groupes de variables.

De même, on a la relation : $V_{XZ} V_{ZZ}^{-1} V_{ZX} a = \beta^2 V_{XX} a$

Un raisonnement analogue à celui fait lors de l'analyse générale nous prouverait de la même façon que les vecteurs propres suivants pris dans l'ordre des valeurs propres décroissantes, correspondent aux combinaisons linéaires de chaque ensemble les plus corrélées entre elles (les combinaisons linéaires relatives à un même ensemble étant assujetties à être non-corrélées).

Remarque 1 : Supposons que la matrice Z n'ait qu'une colonne (q=1) alors, b n'a qu'une composante b, V_{ZZ} est un scalaire (variance de Z) et la relation (C) s'écrit :

$$\beta^2 = \frac{V_{ZX} V_{XX}^{-1} V_{XZ}}{V_{ZZ}}$$

La comparaison avec la formule donnant le coefficient de corrélation multiple nous montre que le coefficient de corrélation canonique β^2 est une généralisation du coefficient R^2, qui se réduit à R^2 dans le cas où l'un des groupes n'est constitué que d'une seule variable.

Remarque 2 : Toujours dans le cas où Z n'a qu'une colonne, la relation précédente :

$$a = \frac{1}{\beta} V_{XX}^{-1} V_{XZ} b$$

nous prouve que le vecteur a est proportionnel, au coefficient b/β près, au vecteur $V_{XX}^{-1} V_{XZ}$ des coefficients de régression (Régression multiple expliquant Z par les p variables du premier groupe).

Le coefficient b/β est d'ailleurs facile à calculer puisque d'après les relations de normalisation, $b = 1/\sqrt{V_{ZZ}}$; β n'est autre que R.

Remarque 3 : Les deux relations :

$$a = \frac{1}{\beta} V_{XX}^{-1} V_{XZ} b \qquad (D)$$

$$b = \frac{1}{\beta} V_{ZZ}^{-1} V_{ZX} a \qquad (E)$$

peuvent s'écrire en revenant aux définitions de V_{XX}, V_{ZZ}, V_{XZ} :

$$a = \frac{1}{\beta} (\tilde{X}\, X)^{-1} \tilde{X}\, Z\, b$$

$$b = \frac{1}{\beta} (\tilde{Z}\, Z)^{-1} \tilde{Z}\, X\, a$$

Prémultipliant les deux membres de chacune d'elles respectivement par X et Z on obtient :

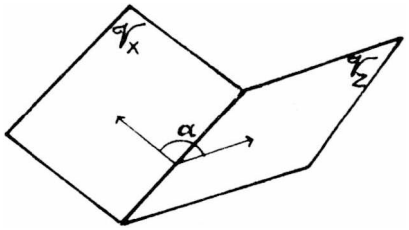
$$Xa = \frac{1}{\beta} X(\tilde{X}\, X)^{-1} \tilde{X}\, Z\, b \qquad (F)$$

$$Zb = \frac{1}{\beta} Z(\tilde{Z}\, Z)^{-1} \tilde{Z}\, X\, a \qquad (G)$$

Désignons par $\mathcal{V}_X$ et $\mathcal{V}_Z$ les variétés linéaires de $\mathbb{R}^{p+q}$ engendrées respectivement par les colonnes de X et de Z.

Les combinaisons linéaires a et b définissent des points de $\mathcal{V}_X$ et de $\mathcal{V}_Z$ qui ont respectivement pour coordonnées Xa et Xb.

Les matrices idempotentes : $Po = X(\tilde{X}\,X)^{-1}\tilde{X}$ et $Q_o = Z(\tilde{Z}\,Z)^{-1}\tilde{Z}$ sont les opérateurs-projection respectivement sur les variétés $\mathcal{V}_X$ et $\mathcal{V}_Z$. Autrement dit les relations (F) et (G) expriment que chacun des vecteurs est projection



de l'autre, ce qui est naturel puisque l'on cherche deux vecteurs Xa et Zb faisant un angle minimal : les vecteurs Xa et Zb étant unitaires, les formules précédentes nous montrent en effet que $\beta = \cos\alpha = \cos(Xa,Zb)$ (la première racine canonique β^2 est le carré du cosinus du plus petit angle entre les sous-espaces $\mathcal{V}_X$ et $\mathcal{V}_Z$).

Notons que ces considérations géométriques nous auraient permis d'écrire directement les formules (F) et (G), et donc de procéder au calcul des variables canoniques (en substituant par exemple dans la relation (G) Xa par sa valeur tirée de la relation (F) sans faire intervenir de multiplicateurs de LAGRANGE.

V - ANALYSE DISCRIMINANTE

On désigne sous le nom d'analyse discriminante toute une série de techniques destinées à décrire et à classer des individus caractérisés par un nombre important de variables. Nous ne ferons que tracer quelques grandes lignes de la principale méthode utilisée pour montrer son lien avec les autres techniques dont il a été question dans ce chapitre.

V-1. FORMULATION DU PROBLEME ET NOTATIONS

Considérons pour fixer les idées le tableau de données qui contient une répartition en 100 postes des dépenses annuelles de 1 000 ménages dont le chef est salarié. Il existe une partition des 1 000 ménages selon 9 catégories socio-professionnelles du chef de ménage.

On peut se poser la question suivante : Etant donné un ménage supplémentaire dont on connaitrait les 100 types de consommations annuelles, peut-on dès lors prévoir sa catégorie socio-professionnelle ?

La question est ici artificielle ; elle peut assouvir une certaine curiosité scientifique, mais ne répond pas à un besoin pratique. Or il peut arriver que la mesure de nombreuses variables sur un individu supplémentaire soit le seul moyen de l'affecter à une classe particulière : il en est parfois ainsi dans le domaine médical où seuls de nombreux examens ou analyses permettent de savoir si un individu a telle ou telle maladie, doit ou ne doit pas être opéré, etc... L'analyse discriminante permettra alors de réaliser un diagnostic ou de fournir des éléments d'information en vue d'une décision particulière.

Soit $X = (x_{ij})$ le tableau des données à n lignes (individus ou observations) et p colonnes (variables).

Les n lignes sont partitionnées en q classes.

On notera : $$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

$\bar{x}_j$ est la moyenne générale de la variable j.

La classe k est caractérisée par un sous-ensemble I_k de n_k valeurs de l'indice i, avec bien entendu

$$\sum_{k=1}^{q} n_k = n$$

$\bar{x}_{kj}$ désigne la moyenne de la variable j pour la classe k :

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} x_{ij}$$

On a la relation, pour toute variable j :

$$\bar{x}_j = \sum_{k=1}^{q} \left(\frac{n_k}{n} \right) . \bar{x}_{kj}$$

La covariance globale de deux variables j et j' s'écrit :

$$\text{cov}(j,j') = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) =$$

$$\frac{1}{n} \sum_{k=1}^{q} \left[\sum_{i \in I_k} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \right]$$

Comme en analyse de la variance, nous allons décomposer cov(j,j') en somme de covariances intra-classes (à l'intérieur des classes) et covariances interclasses (entre les classes).

Pour cela, nous partirons de l'identité, pour tout i, j, k :

$$(x_{ij} - \bar{x}_j) = (x_{ij} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j)$$

La somme entre crochets se décompose alors en quatre termes, dont deux sont nuls.

En effet : $\sum_{i \in I_k} (x_{ij} - \bar{x}_{kj})(\bar{x}_{kj'} - \bar{x}_{j'}) = (\bar{x}_{kj'} - \bar{x}_{j'}) \sum_{i \in I_k} (x_{ij} - \bar{x}_{kj}) = 0$

(par définition de $\bar{x}_{kj}$)

De la même façon :

$$\sum_{i \in I_k} (\bar{x}_{kj} - \bar{x}_j)(x_{ij'} - \bar{x}_{kj'}) = 0$$

Il reste la formule dite de décomposition de HUYGHENS, ou équation d'analyse de la variance :

$$\text{cov}(j,j') = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \in I_k} (x_{ij} - \bar{x}_{kj})(x_{ij'} - x_{kj'}) +$$

$$\sum_{k=1}^{q} \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'})$$

que nous noterons matriciellement :

$$T = D + E \qquad (A)$$

avec $t_{jj'} = \text{cov}(j,j')$

$$d_{jj'} = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \in I_k} (x_{ij} - \bar{x}_{kj})(x_{ij'} - \bar{x}_{kj'})$$

$$e_{jj'} = \sum_{k=1}^{q} \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'})$$

(Mnémotechniquement : la covariance Totale est la somme de la covariance Dans les classes, et de la covariance Entre les classes).

Soit maintenant $u(i)$ la valeur, pour l'individu i, d'une combinaison linéaire des p variables centrées :

$$u(i) = \sum_{j=1}^{p} u_j (x_{ij} - \bar{x}_j)$$

La variance $v(u)$ de la nouvelle variable synthétique $u(i)$ vaut, puisque $u(i)$ est centrée :

$$v(u) = \sum_{i=1}^{n} u^2(i) = \sum_{i=1}^{n} \left[\sum_{j=1}^{p} u_j (x_{ij} - \bar{x}_j) \right]^2$$

$$v(u) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{j'=1}^{p} u_j u_{j'} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

En intervertissant les sommations :

$$v(u) = \sum_{j=1}^{p} \sum_{j'=1}^{p} u_j u_{j'} \mathrm{cov}(j,j') = \tilde{u}\, T u$$

(u désigne le vecteur dont les p composantes sont $u_1, \ldots\ldots u_p$).

Ainsi, la variance d'une combinaison linéaire u de variables se décompose d'après la relation (A) en variance interne et variance externe :

$$\tilde{u}\, T u = \tilde{u}\, D u + \tilde{u}\, E u \qquad (B)$$

Le problème que se propose de résoudre l'analyse discriminante peut alors se formuler ainsi : Parmi toutes les combinaisons linéaires de variables (qui sont, rappelons-le, des opérateurs projection sur des sous-espaces à 1 dimension de $\mathbb{R}^p$), cherchons celles qui ont une variance externe maximale (afin d'exalter les différences entre classes) et une variance interne minimale (afin que l'étendue des classes soit bien délimitée). Ces combinaisons linéaires sont les "fonctions discriminantes".

Il s'agit de chercher u tel que le quotient $\tilde{u}\, Eu / \tilde{u}\, Du$ soit maximal (ou $\tilde{u}\, Du / \tilde{u}\, Eu$ minimal).

Ce qui revient également, d'après la relation (B) à rendre minimal : $\tilde{u}\, Tu / \tilde{u}\, Eu$, ou maximal : $f(u) = \tilde{u}\, Eu / \tilde{u}\, Tu$.

V-2. CALCUL DES FONCTIONS DISCRIMINANTES

La fonction $f(u)$ à maximiser étant homogène de degré 0 en u (invariante si u est changé en αu, α étant un scalaire quelconque), il revient au même de chercher le maximum de $\tilde{u}\, Eu$ avec la contrainte $\tilde{u}\, Tu = 1$

Nous sommes conduits à la relation matricielle :

$$2Eu - 2\lambda Tu = 0$$

Soit $$Eu = \lambda Tu \qquad (C)$$

En général, la matrice des covariances totale T sera inversible, d'où :

$$T^{-1} Eu = \lambda u$$

u est donc vecteur propre de $T^{-1}E$ relatif à la plus grande valeur propre λ. (En effet, en prémultipliant les deux membres de (C) par $\tilde{u}$, on constate que $\tilde{u}\, Eu$, le maximum cherché n'est autre que λ).

La plus grande valeur propre λ, quotient de la variance externe de la fonction discriminante par la variance totale, est inférieure à 1, d'après la relation (B). On l'appelle quelquefois "pouvoir discriminant de la fonction u".

Remarque : La matrice $T^{-1}E$, à p lignes et p colonnes n'est pas symétrique. Il est possible de se ramener à la diagonalisation d'une matrice (q,q) symétrique (rappelons que p est le nombre de variables, et q le nombre de classes).

En effet, la matrice E, de terme général :

$$e_{jj'} = \sum_{k=1}^{q} \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'})$$

est le produit d'une matrice C à p lignes et q colonnes par transposée :

C a pour terme général :

$$c_{jk} = \sqrt{\frac{n_k}{n}}\ (\bar{x}_{kj} - \bar{x}_j) \qquad (D)$$

La relation (C) s'écrit :

$$C\tilde{C}\, u = \lambda Tu$$

Posons $u = T^{-1}Cv$

La relation (C) s'écrit alors :

$$C\tilde{C}\, T^{-1}Cv = \lambda Cv \qquad (E)$$

Il est clair que tout vecteur propre v de la matrice symétrique d'ordre (q,q) : $\tilde{C}\, T^{-1} C$, relatif à une valeur propre λ différente de 0 vérifie également (E). Le vecteur $u = T^{-1}Cv$, et le scalaire λ vérifient alors la relation (C).

Il suffit donc en pratique d'effectuer la diagonalisation de la matrice symétrique $\tilde{C}\, T^{-1} C$, puis d'en déduire u par la transformation $u = T^{-1}Cv$.

V-3. LIEN AVEC L'ANALYSE CANONIQUE

Désignons toujours notre tableau de données à n lignes et p colonnes par X, et notons Y le tableau des variables centrées de terme général : $y_{ij} = x_{ij} - \bar{x}_j$.

Nous allons coder l'information relative à notre partition des n individus en q classes en construisant une matrice Z à n lignes et q colonnes, l'élément Z_{ik} de la k-ème colonne de Z valant 1 si l'individu i appartient à la classe k, 0 dans le cas contraire.

Nous poserons, comme précédemment :

$$D = (Y,Z)$$

Autrement dit, un peu comme en analyse de la covariance, nous ajoutons aux variables initiales des variables artificielles qui indiquent l'appartenance aux diverses classes.

Notons qu'à la différence du paragraphe IV, les colonnes de Z ne sont pas centrées, la somme des éléments de la k-ème colonne de Z vaut n_k, et n'est donc pas nulle.

Explicitons les blocs de la matrice :

$$V(D) = \frac{1}{n} \begin{bmatrix} \tilde{Y} X & \tilde{Y} Z \\ \tilde{Z} Y & \tilde{Z} Z \end{bmatrix}$$

en tenant compte notamment de la nature particulière des colonnes de Z.

La matrice $V_{XX} = \frac{1}{n} \tilde{Y} Y$ n'est autre que la matrice des covariances globale désignée précédemment par la lettre T.

La matrice $V_{ZZ} = \frac{1}{n} \tilde{Z} Z$ est diagonale, et son k-ème élément diagonal vaut n_k/n, effectif relatif de la k-ème classe - (en effet,

$$\sum_{i=1}^{n} Z_{ik} Z_{i\ell} = \delta_{k\ell} \cdot n_k,$$

car l'individu i appartient soit à la classe k, soit à la classe ℓ - Si $k = \ell$, il y aura autant de termes non nuls dans la somme que d'individus dans la classe k).

La matrice à p lignes et q colonnes $V_{XZ} = \frac{1}{n} \tilde{Y} Z$ a pour terme général :

$$(V_{YZ})_{jk} = \frac{1}{n} \sum_{i=1}^{n} y_{ij} Z_{ik} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) Z_{ik} = \frac{1}{n} \sum_{i \in I_k} (x_{ij} - \bar{x}_j)$$

$$\text{Soit} \quad (V_{YZ})_{jk} = \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)$$

La relation (D) nous montre que :

$$(V_{XZ})_{jk} = \sqrt{\frac{n_k}{n}}\, c_{jk}$$

Si nous effectuons formellement l'analyse canonique du tableau partitionné D (nous disons formellement car le sous-tableau Z n'est pas constitué de variables centrées, l'interprétation des calculs est légèrement modifiée),

Nous sommes conduits à chercher un vecteur a satisfaisant l'équation :

$$V_{XZ} V_{ZZ}^{-1} V_{ZX}\ a = \beta^2 V_{XX}\ a \quad \text{(avec } \beta^2 \text{ maximal)}$$

Or on a la relation :

$$V_{XZ} V_{ZZ}^{-1} V_{ZX} = C\tilde{C} = E$$

(En effet :

$$\sum_{k=1}^{q} c_{jk} \sqrt{\frac{k_n}{n}} \left(\frac{n}{n_k}\right) c_{j'k} \sqrt{\frac{n_k}{n}} = \sum_{k=1}^{q} c_{jk} c_{j'k} = e_{jj'})$$

On a vu d'autre part que $V_{XX} = T$

Le vecteur vérifie donc la relation :

$$E\,a = \beta^2\, Ta$$

qui n'est autre que la relation (C)

Nous pouvons également noter que l'on a, pour les deux types d'analyse, la même contrainte de normalisation : $\tilde{a}\, Ta = 1$

Il y a coïncidence entre variable canonique et fonction discriminante.

- L'analyse discriminante apparaît donc comme un cas particulier de l'analyse canonique non centrée, lorsque l'un des deux ensembles est constitué de vecteurs booléens (constitués de 0 ou de 1) décrivant chacun les éléments de la partition de l'ensemble des individus.

V-4. CAS DE DEUX CLASSES

Dans ce cas, rencontré fréquemment lors des applications, de nombreuses simplifications apparaissent. L'analyse discriminante est alors un cas particulier de la régression multiple, où la variable expliquée ne prend que deux modalités, chacune d'elles caractérisant une classe.

La matrice des covariances entre classes a pour terme général :

$$e_{jj'} = \frac{n_1}{n}(\bar{x}_{1j} - \bar{x}_j)(\bar{x}_{1j'} - \bar{x}_{j'}) + \frac{n_2}{n}(\bar{x}_{2j} - \bar{x}_j)(\bar{x}_{2j'} - \bar{x}_{j'})$$

avec :

$$\bar{x}_j = \frac{n_1}{n}\bar{x}_{1j} + \frac{n_2}{n}\bar{x}_{2j}$$

En remplaçant $\bar{x}_j$ par sa valeur, et en tenant compte du fait que

$$n_1 + n_2 = n$$

On trouve sans difficulté :

$$e_{jj'} = \frac{n_1 n_2}{n^2}(\bar{x}_{1j} - \bar{x}_{2j})(\bar{x}_{1j'} - \bar{x}_{2j'})$$

La matrice symétrique d'ordre (p, p) E est ainsi le produit d'une matrice colonne c par sa transposée : $E = c\tilde{c}$

$$c_j = \frac{\sqrt{n_1 n_2}}{n} (\bar{x}_{1j} - \bar{x}_{2j})$$

La relation (C) s'écrit encore :

$$T^{-1} c\tilde{c}\, u = \lambda u$$

Prémultiplions les deux membres par $\tilde{c}$

$$\left[\tilde{c}\, T^{-1} c\right] \tilde{c}\, u = \lambda \tilde{c}\, u$$

La quantité entre crochets est un scalaire, égal par conséquent à λ qui est ici une valeur propre unique.

La valeur propre $\lambda = \tilde{c}\, T^{-1} c$ est appelée "Distance généralisée" des deux classes, et le vecteur propre correspondant, $u = T^{-1} c$ est l'unique fonction discriminante ; (la quantité λ est parfois appelée "D^2 de MAHALANOBIS").

Remarque : Considérons un vecteur y à n composantes, tel que $y_i = \sqrt{n_2/n_1}$ si le i-ème individu appartient à la classe 1, $y_i = -\sqrt{n_1/n_2}$ s'il appartient à la classe 2. La régression multiple expliquant y par les colonnes de X nous donne un vecteur de coefficient $a = V_{XX}^{-1} V_{XY}$ avec ici $V_{XX} = T$, et comme on peut le vérifier aisément : $V_{XY} = c$.

Le vecteur des coefficients de régression a coïncide donc avec le vecteur des composantes de la fonction discriminante u. .

VI - CONTROLE DE VALIDITE DES RESULTATS EN ANALYSE FACTORIELLE

Tests d'hypothèse et simulation

VI-1. LA VALIDITE DES RESULTATS EN STATISTIQUE

Les méthodes d'analyses factorielles dont nous avons parlé jusqu'à présent ont un assez grave inconvénient : elles fournissent toujours un résultat ! Il s'agit d'un inconvénient malheureusement familier en statistique : un simple calcul de moyenne, ou de régression, fournit également toujours un résultat, considéré généralement comme l'estimation d'un paramètre idéal, dont l'existence découle d'hypothèses concernant la population qui est supposée avoir généré les observations.

Les aspects pratiques de la démarche du statisticien sont généralement les suivants : une hypothèse, qui peut être simple ou composite, est faite au sujet de la population parente. Nous désignerons cette hypothèse par H_o.

Une certaine fonction des observations est alors calculée, dont on sait que, sous l'hypothèse H_o, elle suit une certaine loi, préalablement tabulée.

On regarde alors sur la table si la valeur de la fonction calculée correspond à quelque chose de plausible, c'est-à-dire appartient à un intervalle qui contient habituellement 95% (par exemple) des occurrences des valeurs de fonctions du même type, sous les mêmes hypothèses.

Cette démarche, qui a fait ses preuves dans de nombreux domaines d'application (contrôle de fabrication, expérimentation biologique, etc...) n'est pas toujours applicable, ni satisfaisante lorsqu'elle est appliquée en sciences humaines ou en sciences économiques.

En effet, la méthode précédente est parfois inversée, en ce sens que les hypothèses ne sont pas faites d'après la seule considération des données statistiques, de leur contexte réel, mais en fonction de l'existence de modèles théoriques simples, ou de la disponibilité de tables statistiques.

Or une table statistique n'existe que si la loi de la "fonction des obser-

vations" permet des calculs analytiques (cas des tables classiques) ou ne dépend que d'un nombre raisonnable de paramètres (au cas où les tables seraient établies par simulation).

Ces fonctions relativement simples des observations sont forcément très limitées et rarement adéquates : dans le domaine d'application qui nous intéresse plus particulièrement, les observations sont rarement indépendantes, ou ont rarement le même poids, d'où une complication inextricable des éventuels modèles...

Nous nous trouvons bien, en analyse factorielle, dans une situation où la théorie des tests classiques ne peut pleinement et valablement s'appliquer, notamment à cause de la complexité des "fonctions" mises en jeu.

Le problème essentiel est de savoir ce que valent les représentations obtenues dans l'espace des premiers facteurs : il nous faut donc connaître la loi des valeurs propres calculées au cours des analyses, afin de savoir si elles sont "anormalement" élevées, et donc si les facteurs qui leur correspondent extraient bien une part significative de la dispersion totale. Connaître la loi des valeurs propres permet donc de savoir quelle est la dimension "n" du sous-espace de représentation, formée par conséquent des "n" premiers facteurs.

Mais ici, d'une part l'hypothèse H_o sous laquelle on peut calculer la loi des valeurs propres est souvent beaucoup trop restrictive (variables normales indépendantes d'écart-type unité, par exemple), d'autre part, même sous une telle hypothèse, la complexité des résultats obtenus les rend très difficilement utilisables.

Pour fixer les idées, si les variables analysées sont normales, multidimensionnelles, de matrice de covariances théoriques égale à la matrice unité, la densité de probabilité des valeurs propres s'écrit, en fonction du nombre n d'observations et du nombre p de variables :

$$dF = \frac{\pi^{P/2}}{2^{\frac{p(n-1)}{2}}} \prod_{j=1}^{P} \frac{\lambda_j^{\frac{1}{2}(n-p-2)} \exp\left\{-\frac{1}{2}\sum \lambda_j\right\}}{\Gamma\left(\frac{1}{2}(n-j)\right)\Gamma\left(\frac{1}{2}(p+1-j)\right)} \prod_{j<k} (\lambda_j - \lambda_k) \prod_{j<k} d\lambda_j$$

Cette formule est pratiquement inutilisable ; il faut donc procéder

autrement pour savoir combien de facteurs retenir lors des analyses factorielles.

VI-2. LES PROGRAMMES-TESTS

Supposons que l'on veuille savoir si la première valeur propre peut vraisemblablement provenir d'un échantillon E pour lequel l'hypothèse H_o est vérifiée. Il nous suffit de générer quelques échantillons simulés E_i, et de calculer les différentes valeurs propres $f(E_i)$ que l'on comparera à la valeur initiale $f(E)$.

Si n-1 réalisations ont été simulées, et si la loi de $f(E)$ est la même que celle des $f(E_i)$, autrement dit si H_o est vérifiée, alors $f(E)$ a une chance sur n d'être supérieur aux $f(E_i)$.

Si l'on réalise par exemple 19 simulations E_1, E_2, ..., E_{19}, la valeur observée a une chance sur 20 (c'est-à-dire 5 chances sur 100, seuil usuel en statistique) d'être supérieure à l'ensemble des valeurs simulées.

En pratique, pour des analyses de tableaux particulièrement importants, on pourra se limiter à un nombre beaucoup plus restreint de simulations, en instituant des procédures d'arrêt (ceci afin d'éviter une exploitation qui peut être coûteuse en temps-machine).

1) - Arrêt si la valeur observée est dépassée par une valeur simulée.

2) - Arrêt si la valeur observée est bien plus grande que, par exemple, les 5 premières valeurs simulées, à l'aide d'une sorte de "t" de STUDENT, ou de toute autre fonction permettant d'apprécier la distance d'une observation à un petit échantillon.

Il reste donc, pour chaque type de problème, à choisir une hypothèse H_o convenable, et à générer des échantillons sous cette hypothèse.

VI-3. REALISATIONS PRATIQUES

Les simulations seront adaptées à chaque type de données, en fonction des hypothèses H_o proposées par les utilisateurs eux-mêmes.

1) - Tableaux de valeurs numériques (n individus, p variables)

La méthode la plus simple consiste à permuter aléatoirement les n observations initiales correspondant à chacune des p variables.

L'amplitude des valeurs propres issues de l'analyse de ces tableaux correspond bien à un "bruit de fond" que l'utilisateur veut éliminer.

2) - Tableaux de contingences (n lignes, p colonnes)

Les cases de ces tableaux correspondent à des effectifs x_{ij} (ou des fréquences p_{ij}). On simule alors une correspondance aléatoire par le schéma multinomial suivant : si $x..$ désigne l'effectif total :

$$x.. = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}$$

La variable simulée X_{ij} est une variable aléatoire normale (approximation classique de la loi multinomiale) de moyenne

$$E(X_{ij}) = x..\ f_{i.}\ f_{.j}$$

et de variance $\quad var\ (X_{ij}) = x..\ f_{i.}\ f_{.j}(1 - f_{i.}\ f_{.j})$

Les paramètres $x..$, $f_{i.}$, $f_{.j}$ correspondent aux valeurs observées initialement.

REFERENCES BIBLIOGRAPHIQUES SOMMAIRES

DU CHAPITRE I

1 - ANDERSON T.W. - An introduction to multivariate analysis -
WILEY and Son - New-York - 1958

2 - BENZECRI J.P. - L'analyse des données - Tome 2

L'analyse des correspondances - DUNOD - 1973

3 - LEBART L. et FENELON J.P. - Statistique et Informatique Appliquées
DUNOD - 2ème édition - 1973.

CHAPITRE II

ANALYSE DE CERTAINES CORRESPONDANCES MULTIPLES

I - GENERALITES

Une partie généralement importante des fichiers d'enquête se compose de réponses à des questions mises sous forme disjonctive complète, c'est-à-dire de questions dont les diverses modalités de réponses s'excluent mutuellement, et telles qu'une modalité est obligatoirement choisie.

L'ensemble des r modalités de réponses à une telle question permet de partitionner l'échantillon en r classes, au plus.

Exemple 1 : Intitulé de la question : âge du père

8 modalités - 1°/ moins de 25 ans
2°/ de 25 à 29 ans
3°/ de 30 à 34 ans
4°/ de 35 à 39 ans
5°/ de 40 à 44 ans
6°/ de 45 à 49 ans
7°/ 50 ans et plus
8°/ Sans objet ou non-réponse.

Exemple 2 : Intitulé de la question : "Avez-vous un (ou plusieurs) lave-vaisselle ?" :

2 modalités - 1/ oui
2/ non

La donnée de deux questions mises sous forme disjonctive complète nous permet d'observer deux partitions de l'ensemble des individus enquêtés. L'analyse du tableau de correspondance croisant ces deux partitions peut être généralisée au cas de Q partitions (Q étant un entier supérieur à 2). La généralisation que nous proposons (qui n'est évidemment pas la seule possible) est assez naturelle et conduit à des règles d'interprétation simples.

L'exposé élémentaire ci-dessous est complété par les listages des programmes (conçus pour l'exploitation de fichiers volumineux) et un exemple d'application.

Un exposé plus théorique de cette question figure dans la note de J.P. BENZECRI "Sur l'analyse des tableaux binaires associés à une correspondance multiple", à laquelle nous emprunterons notamment certaines notations.

Notations : L'ensemble des questions sera désigné par Q. Une question q consiste en un ensemble J_q de card J_q modalités.

$J = \cup\{J_q \mid q \in Q\}$ désigne la réunion de tous ces ensembles de modalités.

On désignera par $I = \pi\{J_q \mid q \in Q\}$ l'ensemble produit des J_q, c'est-à-dire l'ensemble dont les éléments sont constitués des suites de q modalités, chacune de celles-ci étant prise dans une question différente. Les éléments de I sont donc les réponses possibles des sujets enquêtés.

L'ensemble des individus enquêtés sera désigné par S.

Contrairement aux approches classiques des tables de contingence multiple, où l'on s'intéresse principalement aux fréquences $k(i)$ des individus ayant donné la réponse i $(i \in I)$, l'étude qui va suivre reste intéressante si card S est très inférieure à card I, ce qui est souvent le cas dans les applications. Si l'on pose à 1000 individus 12 questions ayant chacune 10 modalités : card $S = 10^3$, card $I = 10^{12}$.

Ainsi dans le tableau $k(i)$, un millionième seulement des éléments sont différents de 0.

On désignera par Z le tableau (card S x card J) donnant, pour l'individu $s \in S$ une description booléenne de ses réponses aux card Q questions.

Si $\varrho(s,q)$ désigne la modalité de la question q choisie par le sujet s, $(\varrho(s,q) \in J_q)$, et si $q(j)$ est l'indice q de la question à laquelle appartient la modalité $j \in J$, le terme générique de Z s'écrit :

$$z_{sj} = \delta^{j}_{\varrho(s,q(j))}$$

Le tableau des éléments $\varrho(s,q)$ constitue un codage condensé du tableau Z. (Le tableau de terme général $\varrho(s,q)$ n'a que card Q colonnes).

Les programmes donnés ci-après n'utilisent que ce type de tableaux comme entrée. A l'intérieur de chaque question J_q, les modalités sont indicées de 1 à card J_q. $\varrho(s,q)$ n'est autre que la valeur de cet indice, pour l'individu s et la question q.

II - TABLEAU DE BURT ASSOCIE A Z

Partitionnons les colonnes de la matrice Z de façon à faire apparaître dans un même sous-tableau Z_q les colonnes relatives aux modalités de la question q.

Dans ces conditions : $Z = (Z_1, Z_2, \ldots\ldots Z_{\text{card } Q})$.

Si $\tilde{Z}$ désigne la transposée de Z, le tableau :

$$B = \tilde{Z}Z$$

est appelé "Tableau de contingence de BURT" associé au tableau des réponses Z.

Le tableau B est formé de $(\text{card } Q)^2$ blocs.

Le q-ème bloc diagonal $\tilde{Z}_q Z_q$ est une matrice diagonale d'ordre $(\text{card } J_q)^2$. (Puisque deux modalités d'une même question ne peuvent être choisies simultanément).

Le bloc indicé par (q,q'), d'ordre (card J_q x card $J_{q'}$) n'est autre que le tableau de contingence croisant les réponses aux deux questions q et q'.

Nous désignerons par T, d'ordre (card J x card J), la matrice diagonale ayant les mêmes éléments diagonaux que B (ces éléments diagonaux ne sont autres que les effectifs correspondant à chacune des modalités).

La matrice T peut être également considérée comme formée de $(\text{card } Q)^2$ blocs (seules les card Q blocs diagonaux sont des matrices non nulles, le

q-ème bloc diagonal T_q valant $\tilde{Z}_q Z_q$, matrice dont les termes diagonaux sont les effectifs correspondant aux diverses modalités de la question q).

Cas de deux questions : (card Q = 2).

le tableau des réponses Z s'écrit alors : $Z = (Z_1, Z_2)$.

Il est alors équivalent, du point de vue de la description des associations entre modalités :

1/ d'effectuer l'analyse des correspondances du tableau Z d'ordre (card S, card J).
2/ d'effectuer l'analyse des correspondances du tableau B d'ordre (card J x card J).
3/ d'effectuer l'analyse des correspondances du tableau $\tilde{Z}_1 Z_2$ d'ordre (card J_1 x card J_2).

Montrons que les analyses (1) et (2) fournissent les mêmes facteurs (à une normalisation près).

Les facteurs issus de (1) vérifient l'équation :

$$\frac{1}{\text{card}Q}\, T^{-1}\, \tilde{Z}\, Z\, \varphi = \lambda \varphi \qquad (1)$$

D'autre part, les marges du tableau B sont les éléments diagonaux de la matrice card Q.T

La relation de transition relative à l'analyse de B s'écrit, puisque $B = \tilde{Z}Z$ est symétrique (et définie non négative), pour un facteur ψ relatif à la valeur propre μ :

$$\frac{1}{\text{card}Q}\, T^{-1}\, \tilde{Z}\, Z\, \psi = \sqrt{\mu}\, \psi \qquad (2)$$

Ainsi, $\psi = \varphi$ et $\mu = \lambda^2$

Montrons maintenant que pour tout couple de facteurs (φ_1, φ_2) associé à la même valeur propre β lors de l'analyse du tableau de contingence $\tilde{Z}_1 Z_2$ correspond un facteur $\varphi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}$ de l'analyse de B (ou de Z).

Les deux marges du tableau rectangulaire $\tilde{Z}_1 Z_2$ sont T_1 et T_2.

Les deux relations de transition s'écrivent :

$$T_1^{-1}\tilde{Z}_1 Z_2\ \varphi_2 = \sqrt{\beta}\,\varphi_1 \qquad (3)$$

$$T_2^{-1}\tilde{Z}_2 Z_1\ \varphi_1 = \sqrt{\beta}\,\varphi_2 \qquad (4)$$

Ou encore :

$$T_1^{-1}(T_1\,\varphi_1 + \tilde{Z}_1 Z_2\ \varphi_2) = (1+\sqrt{\beta})\ \varphi_1$$

$$T_2^{-1}(T_2\,\varphi_2 + \tilde{Z}_2 Z_1\ \varphi_1) = (1+\sqrt{\beta})\ \varphi_2$$

qui s'écrit également, après multiplication des deux membres par $1/2 = 1/\text{cardQ}$

$$\frac{1}{2}.T^{-1}\tilde{Z}Z\ \varphi = \frac{(1+\sqrt{\beta})}{2}\ \varphi$$

Ce qui n'est autre que la relation (1) où $\lambda = \dfrac{1+\sqrt{\beta}}{2}$

Ainsi, les valeurs propres issues des trois analyses sont respectivement $\lambda,\ \lambda^2,\ (2\lambda - 1)^2$.

III - GENERALISATION AU CAS DE PLUS DE DEUX QUESTIONS

Le tableau $Z = (Z_1, Z_2, \ldots Z_q \ldots Z_{\text{cardQ}})$ possède cardJ colonnes, auxquelles correspondent cardJ points de $\mathbb{R}^{\text{cardS}}$; plaçons-nous dans l'espace $\mathbb{R}^{\text{cardS}}$. Le sous-tableau Z_q engendre une variété linéaire V_q à card J_q dimensions.

Toutes ces variétés linéaires ont au moins en commun la première bissectrice. Le rang du tableau Z est donc au plus égal à cardJ -(cardQ - 1).

Soit φ_q le tableau (à cardJ_q lignes et une colonne) des composantes d'un point $M_{(q)}$ de V_q dans la base définie par les colonnes de Z_q.

Le carré de distance de ce point $M_{(q)}$ à l'origine, selon la norme euclidienne usuelle n'est autre que :

$$\tilde{\varphi}_q \tilde{Z}_q Z_q \varphi_q = \tilde{\varphi}_q T_q \varphi_q$$

L'analyse des correspondances du tableau de contingence croisant deux questions q et q' revient à étudier les positions respectives des variétés $\mathcal{V}_q$ et $\mathcal{V}_{q'}$. En effet, dans $\mathbb{R}^{\text{cardS}}$, les opérateurs-projection sur $\mathcal{V}_q$ et $\mathcal{V}_{q'}$ correspondent aux matrices $Z_q(\tilde{Z}_q Z_q)^{-1}\tilde{Z}_q$ (c'est-à-dire $Z_q T_q^{-1}\tilde{Z}_q$) et $Z_{q'}(\tilde{Z}_{q'}Z_{q'})^{-1}\tilde{Z}_{q'}$, $(= Z_{q'}T_{q'}^{-1}\tilde{Z}_{q'})$.

Les relations de transition (3) et (4) (où $q=1$, $q'=2$) expriment que les points M_q et $M_{q'}$ sont projections l'un de l'autre. Il revient au même de chercher deux points M_q et $M_{q'}$ tels que leur moyenne des carrés des distances à l'origine soit constante.

$$\tilde{\varphi}_q T_q \varphi_q + \tilde{\varphi}_{q'} T_{q'} \varphi_{q'} = 2\ \text{cardS} \qquad (5)$$

et tels que la distance à l'origine du point $M = M_q + M_{q'}$, soit maximale.

$$\|OM\|^2 = \tilde{\varphi}_q T_q \varphi_q + \tilde{\varphi}_{q'} T_{q'} \varphi_{q'} + 2\,\tilde{\varphi}_q \tilde{Z}_q Z_{q'} \varphi_{q'} \qquad (6)$$

$$\|OM\|^2 = 2\ \text{cardS}\left(1 + \left[\frac{1}{\text{cardS}}\tilde{\varphi}_q\ \tilde{Z}_q Z_{q'}\ \varphi_{q'}\right]\right) \qquad (7)$$

Remarque : Le maximum de $\|OM\|^2$ s'obtient d'ailleurs (et ceci ne sera valable que pour deux questions) pour $\tilde{\varphi}_q T_q \varphi_q = \tilde{\varphi}_{q'} T_{q'} \varphi_{q'}$. L'expression entre crochets dans le membre de droite de la relation (7) n'est autre que le cosinus de l'angle des vecteurs $(OM_q, OM_{q'})$.

Posé sous cette dernière forme, le problème se généralise aisément au cas de plus de deux questions.

Si $\varphi_1, \varphi_2, \ldots \varphi_{\text{cardQ}}$ désignent respectivement les vecteurs des composantes de cardQ points $M_1, M_2, \ldots M_{\text{cardQ}}$ dans les bases $Z_1, Z_2, \ldots Z_{\text{cardQ}}$, on cherchera à rendre maximale la quantité :

$$\|OM\|^2 = \sum\{\tilde{\varphi}_q \tilde{Z}_q Z_{q'} \varphi_{q'} \mid q \in Q,\ q' \in Q\} \qquad (5\text{bis})$$

avec la contrainte :

$$\sum\{\tilde{\varphi}_q T_q \varphi_q \mid q \in Q\} = \text{cardQ cardS} \qquad (6\text{bis})$$

Si φ désigne le vecteur à cardJ composantes tel que

$$\tilde{\varphi} = (\tilde{\varphi}_1, \tilde{\varphi}_2, \ldots\ldots, \tilde{\varphi}_{cardQ}) \text{ (ou encore } \varphi = \bigoplus \{\varphi_q \mid q \in Q\})$$

le problème revient à rendre maximum $\tilde{\varphi} B \varphi$ avec $\tilde{\varphi} T \varphi = 1$

Les facteurs φ cherchés sont donc les vecteurs propres de $T^{-1}B$ relatifs aux plus grandes valeurs propres qui sont proportionnels à ceux issus de l'analyse des correspondances du tableau Z (qui coïncident, à une normalisation près, avec ceux issus de l'analyse du tableau B).

IV - PROPRIETES DES ANALYSES MULTIPLES

Les facteurs φ issus de l'analyse du tableau Z vérifient l'équation

$$\frac{1}{cardQ} T^{-1} B \varphi = \lambda \varphi \qquad (8)$$

en faisant apparaître les composantes φ_q de φ relatives à la question q' et les blocs des tableaux T et B, cette équation s'écrit :

$$\frac{1}{cardQ} \sum \{ T_{q'}^{-1} \cdot \tilde{Z}_{q'} Z_q \varphi_q \mid q \in Q \} = \lambda \varphi_{q'}$$

a) - Le centre de gravité du sous-nuage de $cardJ_q$ points dont les coordonnées dans $\mathbb{R}^{cardS}$ sont les colonnes du tableau $Z_q T_q^{-1}$ décrivant les profils des réponses à la question q est le même que le centre de gravité en général du nuage.

Il s'ensuit que les composantes de φ_q relatives à une question particulière q sont également centrées. (Chaque point-modalité j est muni d'une masse égale à $t_{jj}/cardQ.cardS$).

b) - La somme des valeurs propres non triviales vaut, d'après la relation (8) cardJ/cardQ - 1.
(La trace sera donc égale à 1 dans le cas des questions à deux modalités, pour lesquelles cardJ = 2cardQ).

c) - Le carré de distance au centre de gravité d'un point-modalité j ($j \in J$) de $\mathbb{R}^{cardS}$ s'écrit :

$$d^2(0,j) = \sum \{ cardS \ (Z_{ij}/t_{jj} - 1/cardS)^2 \mid i \in S \}$$

Soit, compte tenu de la relation : $\sum \{ Z_{ij} \mid i \in S \} = t_{jj}$

$$d^2(0,j) = cardS \ (1/t_{jj} - 1/cardS)$$

La contribution à l'inertie totale de la modalité j vaut donc

$$c(j) = \left(\frac{t_{jj}}{cardQcardS} \right) d^2(0,j) = \frac{1}{cardQ} \left(1 - \frac{t_{jj}}{cardS} \right)$$

La contribution de la question q à l'inertie totale vaut :

$$C(q) = \sum \{ c(j) \mid j \in J_q \} = \frac{1}{cardQ} \ (card \ J_q - 1)$$

(On vérifie que $\sum \{ C(q) \mid q \in Q \} = card \ J/card \ Q - 1$)

La forme de $c(j)$ nous prouve que la contribution d'une modalité est d'autant plus forte que l'effectif correspondant est plus faible, sans toutefois pouvoir dépasser $1/card \ Q$.

d) - Réduction des dimensions du tableau analysé

Dans l'espace R^{cardS}, les points représentatifs des cardJ modalités ont pour coordonnées les colonnes de ZT^{-1} ;

Nous avons vu que le rang de Z (donc de ZT^{-1}) est au plus égal à cardJ - card Q+1 ; la variété linéaire engendrée par les colonnes de ZT^{-1} contient la première bissectrice. Comme le nuage est dans l'hyperplan T^{-1} orthogonal à la première bissectrice, le nombre de valeurs propres nulles lors de l'analyse du nuage par rapport à son centre de gravité sera de CardQ.

En faisant choix d'une base dans le support du nuage, on se ramènera donc à la diagonalisation d'une matrice symétrique d'ordre (CardJ - CardQ) x (CardJ - CardQ).

C'est ce que fait le programme SBURT dont le listage figure à la fin du chapitre.

e) - Cas particulier : questions à deux modalités.

Dans ce cas, bien que le programme général SBURT puisse s'appliquer sans perte notable de temps, on obtient directement la matrice à diagonaliser, symétrique, qui n'est autre que la matrice des corrélations entre variables, celles-ci n'étant représentées que par une seule de leurs modalités - (CardJ - CardQ = 1/2 CardJ).

Explicitons la relation (8) ci-dessus, où, rappelons-le, T désigne la matrice diagonale ayant les mêmes éléments diagonaux que B.

$$\frac{1}{\text{cardQ}} \sum_{j \in J} \frac{b_{ij}}{b_{ii}} \varphi^j = \lambda \varphi^i \qquad (8bis)$$

L'ensemble J des questions va maintenant être partitionné en deux sous-ensembles de mêmes cardinalités J^1 et J^2, formés respectivement des premières et des deuxièmes modalités de chacune des cardQ questions.

$$J = J^1 \cup J^2 \qquad (J_q = \{j_q^1, j_q^2\} \mid j_q^1 \in J^1,\ j_q^2 \in J^2,\ q \in Q)$$

Notons les relations, pour tout $q \in Q$:

$$\begin{cases} b_{ij_q^1} + b_{ij_q^2} = b_{ii} \\ b_{j_q^1 j_q^1} + b_{j_q^2 j_q^2} = \text{Card S} \end{cases} \qquad b_{j_q^1 j_q^1} \cdot \varphi_{j_q^1} = -b_{j_q^2 j_q^2} \varphi_{j_q^2}$$

Il suffit donc de restreindre la sommation de la relation (8bis) au seul ensemble J^1, dont l'élément courant sera désormais noté j :

$$\frac{1}{\text{cardQ}b_{ii}} \sum \left\{ \left(b_{ij} \varphi^j - \frac{(b_{ii} - b_{ij})\, b_{jj} \varphi_j}{(\text{cardS} - b_{jj})} \right) \mid j \in J^{-1} \right\} = \lambda \varphi^i$$

Ce qui peut s'écrire : $$\sum \left\{ \frac{\text{CardS} \cdot b_{ij} - b_{ii} \cdot b_{jj}}{\text{CardQ}(\text{cardS} - b_{jj}) b_{ii}} \varphi^j \mid j \in J^1 \right\} = \lambda \varphi^i \qquad (9)$$

et de calculer les moments empiriques centrés du second ordre des cardQ variables caractérisées par leurs premières modalités.

$$\text{cov}(i,j) = \frac{1}{\text{cardS}} \left(b_{ij} - \frac{b_{ii}\, b_{jj}}{\text{card S}} \right)$$

$$\text{var}(j) = \frac{1}{\text{cardS}} \left(b_{jj} - \frac{b_{jj}^2}{\text{cardS}} \right)$$

Le terme général de la matrice des corrélations des cardQ variables s'écrit :

$$\text{cor}(i,j) = \frac{\text{Card S}.b_{ij} - b_{ii}b_{jj}}{\left[(\text{CardS}-b_{jj})b_{jj}(\text{CardS}-b_{ii})b_{ii}\right]^{\frac{1}{2}}}$$

Il est clair que si (φ, λ) est solution de l'équation (9) alors (ψ, λ') est solution de :

$$\sum \left\{ \text{cor}(i,j)\ \psi^j \;\middle|\; j \in J_1 \right\} = \lambda'\, \psi^j$$

avec :
$$\begin{cases} \psi^j = \varphi^j . \left[\dfrac{(\text{CardS} - b_{jj})}{b_{jj}} \right]^{\frac{1}{2}} \\ \lambda' = \lambda . \text{ cardQ} \end{cases}$$

f) - Cas où l'analyse d'une correspondance multiple se ramène à celle d'une correspondance binaire.

Le cas d'une correspondance binaire s'est revelé être particulièrement intéressant du point de vue des calculs à mettre en oeuvre, car l'analyse du tableau de BURT d'ordre (CardJ x CardJ) équivalait à l'analyse des correspondances du tableau de contingence croisant les modalités des deux questions, ce qui conduit à diagonaliser une matrice d'ordre (Inf(cardJ_1, cardJ_2)).

Ce résultat peut être généralisé de diverses façons sous certaines conditions (cf. J.P. BENZECRI, op. cité).

Nous retiendrons la propriété suivante, utile pour les applications :

- Si l'ensemble Q des questions est partitionné en deux sous-ensembles Q_1 et Q_2 à l'intérieur desquels les questions sont indépendantes, l'analyse des cardQ questions se réduit à celle d'une correspondance binaire, et donc à la diagonalisation d'une matrice d'ordre Inf(cardJ^1, cardJ^2) où $J^i = \left\{ \cup J_q \;\middle|\; q \in Q_i \right\}$

(Nous dirons ici que deux questions q et q' sont indépendantes si le tableau $\tilde{Z}_q Z_{q'}$ est égal à $1/\text{cardS}.t_q \otimes t_{q'}$, où les vecteurs t_q et $t_{q'}$ ont respectivement pour composantes les éléments diagonaux de $\tilde{Z}_q Z_q$ et $\tilde{Z}_{q'} Z_{q'}$, qui sont également les éléments diagonaux de T_q et $T_{q'}$, de par la définition de ces deux matrices ; en notation matricielle $t_q \otimes t_{q'} = t_q . \tilde{t}_{q'}$)

Ecrivons de nouveau la relation (8) en partitionnant φ en deux blocs φ_{Q1} et φ_{Q2} ($\varphi_{Qi} = \oplus \left\{ \varphi_q \;\middle|\; q \in Q_i \right\}$) et les matrices B et T en quatre blocs, de façon à faire apparaître la dichotomie de $Q = Q_1 \cup Q_2$

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \qquad T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}$$

D'où les deux relations :

$$\begin{cases} \dfrac{1}{\text{cardQ}} \left(T_1^{-1} B_{11} \varphi_{Q1} + T_1^{-1} B_{12} \varphi_{Q2} \right) = \lambda \varphi_{Q1} \\ \dfrac{1}{\text{cardQ}} \left(T_2^{-1} B_{21} \varphi_{Q1} + T_2^{-1} B_{22} \varphi_{Q2} \right) = \lambda \varphi_{Q2} \end{cases}$$

Remarquons que les card Q_1 (resp : card Q_2) blocs diagonaux de $T_1^{-1} B_{11}$ (resp : $T_2^{-1} B_{22}$) sont des matrices unité dont les ordres correspondent aux cardinaux de chacune des questions :

$$\left(q \in Q_i,\ q' \in Q_i,\ q = q' \Longrightarrow T_q^{-1} \tilde{Z}_q Z_q = I_{\text{card } q} \;\middle|\; i \in \{1,2\} \right)$$

On a d'autre part, pour $i \in \{1,2\}$

$$q \in Q_i,\ q' \in Q_i,\ q \neq q' \Longrightarrow T_q^{-1} \tilde{Z}_q Z_{q'} = \frac{1}{\text{cardS}} T_q^{-1} t_q . \tilde{t}_{q'}$$

En désignant par $1_{\text{card } q}$ un vecteur dont les card q composantes valent 1,

$$T_q^{-1} \tilde{Z}_q Z_{q'} = \frac{1}{\text{card } S} \cdot 1_{\text{card } q} \cdot \tilde{t}_{q'}$$

Les relations $\tilde{t}_{q'} \varphi_{q'} = 0$ impliquent finalement que, pour $i \in \{1,2\}$:

$$T_i^{-1} B_{ii} \varphi_{Qi} = \varphi_{Qi}$$

Le système ci-dessus s'écrit alors :

$$\begin{cases} T_1^{-1} B_{12} \varphi_{Q2} = (\lambda \text{ card } Q - 1) \varphi_{Q1} \\ T_2^{-1} B_{21} \varphi_{Q1} = (\lambda \text{ card } Q - 1) \varphi_{Q2} \end{cases}$$

D'où, par substitution :

$$T_2^{-1} B_{21} T_1^{-1} B_{12} \varphi_{Q2} = (\lambda \text{ card } Q - 1)^2 \varphi_{Q2}$$

Ainsi, φ_{Q2} est obtenu par diagonalisation d'une matrice d'ordre (card Q_2). On en déduit facilement φ_{Q1}.

Nous avons en fait implicitement supposé que card $Q_2 \leq$ card Q_1, en choisissant de calculer φ_{Q2} avant φ_{Q1}.

Remarquons que B_{12} est obtenu par juxtaposition des tableaux de contingence croisant l'ensemble des modalités des questions du premier groupe et celles relatives au second groupe. Les marges du tableau B_{12} sont les éléments diagonaux de $\text{card}Q_2 . T_1$ et $\text{card}Q_1 . T_2$.

Les facteurs issus de l'analyse du tableau B_{12} vérifient la relation :

$$\frac{1}{\text{card}Q_1 \text{card}Q_2} T_2^{-1} B_{21} T_1^{-1} B_{12} \psi = \mu \psi$$

Ils sont donc proportionnels aux facteurs trouvés précédemment.

g) - Calcul d'une partition moyenne (Résultats empiriques)

Nous avons vu qu'il existe un nombre card I de réponses possibles au questionnaire. En "agrégeant" les individus ayant des réponses identiques, on obtient un tableau dont la plus grande dimension est inférieure ou égale à card I (dont l'analyse fournit les mêmes résultats que l'analyse de Z).

Cependant, pour le type d'enquête qui nous intéresse plus particulièrement, on a en général card S $\ll$ card I, et la partition de tous les individus selon les divers types de réponses possibles coïncide souvent avec la partition triviale en card S classes.

La typologie des individus observable dans le sous-espace des premiers facteurs n'utilise en fait qu'une partie de l'information fournie par l'ensemble des distances entre individus.

On peut chercher à exhiber une partition J_m des individus, dite partition moyenne des card Q partitions données, le terme moyenne étant entendu au sens (intuitif) suivant : la partie stable de la typologie issue de l'analyse de Z (tableau des réponses d'ordre card S x card J) coïncide avec la partie stable de la typologie issue de l'analyse du tableau d'ordre (card I_m x card J) obtenu en agrégeant les lignes de Z appartenant à une même classe de la partition J_m.

Les classes d'une telle partition constituent un base orthogonale (à composantes booléennes) d'un certain sous-espace à card J_m dimensions de R^{cardS}

Le problème pratique est de savoir si, card J_m étant inférieur à un nombre fixé, une telle base peut contenir le sous-espace des premiers facteurs.

Les résultats empiriques dont nous disposons sont encourageants et permettent de penser que n'importe quelle analyse de fichier peut en fait être obtenue à partir d'une partition moyenne J_m avec card $J_m < 50$ (et donc ne nécessitent qu'une diagonalisation de matrice 50 x 50).

Pour l'exemple d'application du §VI ci-après, une partition en 16 classes suffit à reconstituer de façon satisfaisante le plan des deux premiers facteurs. De plus comme cela est souvent le cas dans ces procé-

dures d'agrégation, une partition moyenne enrichit l'interprétation des résultats.

Nous adjoindrons donc ci-après un programme de partition adapté (et simplifié) pour tenir compte de la structure particulière de ces types de questionnaires. Il s'agit d'une technique d'agrégation autour de centres mobiles, apparentée à la méthode des nuées dynamiques proposée par E. DIDAY.

(Pour la recherche d'une partition moyenne, il semble en effet judicieux de considérer comme centre de classe le point représentatif du profil moyen de la classe).

V - PROGRAMMES D'APPLICATION

Caractéristiques générales.

Les programmes ont été écrits en FØRTRAN IV pour CDC 6600.

Le fichier à analyser est sur bande, et comporte en principe un nombre arbitrairement grand d'individus. Chaque enregistrement logique représente la description d'un individu sous la forme condensée signalée au §1. Cependant, pour les rajouts de variables supplémentaires, nous avons (sous-programmes GUSB et PLUSB) rassemblé les coordonnées des individus sur les premiers axes factoriels dans un tableau G.

Une modification minime permettra de s'affranchir de cet encombrement mémoire, qui peut être gênant dans le cas d'une très grosse enquête et d'un petit ordinateur.

Les variables analysées sont celles situées en tête d'enregistrement, et les variables rajoutées (variables illustratives) sont celles qui suivent. Toutefois, le sous-programme PERMU permet de modifier l'ordre des variables, et rectifie les identificateurs des modalités en conséquence.

Les données de base sont donc :

1/ - le fichier condensé de NTOU variables, dont les NVAR premières seront analysées.
2/ - le tableau NMOD(K) donnant le nombre de modalités de la variable K
3/ - le tableau JCAR(I) des identificateurs en A3 des modalités.
4/ - éventuellement des tableaux NOUV(L) gérant une permutation des variables.

Les ordres d'appel seront classiquement de deux types (à partir du programme principal) :

1 - Analyse directe

1/ PERMU (éventuellement)
2/ DEPB (calcule le tableau de BURT)
3/ IMPI (imprime ce tableau)
4/ SBURT (analyse réduite de ce tableau) qui appelle :
HILB, EIGNVA, ICONT

5/ GUSB (coordonnées des individus)

6/ PLUSB (rajouts de variables supplémentaires)

2 - Analyse après calcul d'une partition moyenne

1/ PARTI (calcul d'une partition moyenne en K classes)

appelle : GROUP (qui appelle ETAL, DIST)
ETAL, JMPR, TIRA

(Le sous-programme ETAL a pour effet d'étaler le codage condensé en codage binaire).

2/ Analyse des correspondances usuelles du tableau TAB issu de PARTI.

```
      SUBROUTINE DEPB(VEC,JCAR,NVAR,NSB,NIND,NMOD,NV,NTOU,T,X)
C***** CALCUL D UN TABLEAU DE DEPENDANCE DE BURT SUR NIND INDIVIDUS,
C***** CARACTERISES PAR NVAR VARIABLES,REPRESENTANT EN TOUT NSB MODAL.
C***** NTOU EST LE NOMBRE TOTAL DE VARIABLES SUR LA BANDE
C***** NV EST UN MAJORANT DE NSB
      DIMENSION VEC(NV,NV)
C***** JCAR IDENTIFIE LES NSB MODALITES
C***** NMOD(K) EST LE NOMBRE DE MODALITES DE LA VARIABLE  K
      DIMENSION JCAR(1),NMOD(1)
      DIMENSION T(1),X(1)
      COMMON/ENSOR/NINT,NSORT
      COMMON/BBB/NB
      DO 10 J=1,NSB
      DO 10 JPRIM=1,NSB
   10 VEC(J,JPRIM)=0
      DO 3 I=1,NIND
      READ(NB)(T(K),K=1,NTOU)
C***** CODAGE BINAIRE DES VARIABLES MISES SOUS FORMES REDUITES
      JDEB=0
      DO 5 J=1,NSB
    5 X(J)=0
      DO 4 K=1,NVAR
      L=JDEB+T(K)+0.000001
C***** PROTECTION
      IF(L.GT.NSB)  GO TO 100
      X(L)=1
      JDEB=JDEB+NMOD(K)
    4 CONTINUE
      DO 6 J=1,NSB
      DO 7 JPRIM=1,J
    7 VEC(J,JPRIM)=VEC(J,JPRIM)+X(J)*X(JPRIM)
    6 CONTINUE
    3 CONTINUE
      DO 308 J=1,NSB
      DO 308 JPRIM=J,NSB
  308 VEC(J,JPRIM)=VEC(JPRIM,J)
      GO TO 102
  100 WRITE(NSORT,101)L,NSB,I,K
  101 FORMAT( 10H ERREUR       I4,10HDEPASSE     I4,10HPOUR I=     I4,
     1  10HET POUR K=   I4,///)
      NSB=0
  102 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE SBURT(X,VEC,EIG,F1,F2,F3,NVAR,NIND,NSB,NV,JCAR,A,NMOD)
C***** CE SOUS PROGRAMME EFFECTUE L ANALYSE D UN TABLEAU DE BURT X ,
C***** CORRESPONDANT A L ENSEMBLE DES NSB MODALITES  DE NVAR VARIABLES,
C*****  EN DIAGONALISANT UNE MATRICE D ORDRE (NSB-NVAR)*(NSB-NVAR)
      DIMENSION A(NV,NV),NMOD(1),NCUM(100)
C***** NMOD(K) EST LE NOMBRE DE MODALITES DE LA VARIABLE K
C***** JCAR IDENTIFIE L ENSEMBLE DES MODALITES
C*****  NV EST UN MAJORANT DE NSB
      DIMENSION X(NV,NV),VEC(NV,NV)
C***** LE TABLEAU X EST DETRUIT .
      DIMENSION DISTO(131),PC(131),PL(131)
      DIMENSION JCAR(1),EIG(1),F1(1),F2(1),F3(1)
      COMMON/ENSOR/NINT,NSORT
      SOM=NIND*NVAR
      DO 1 J=1,NSB
    1 PL(J)=X(J,J)
      PC=NVAR
C***** CALCUL DE LA PREMIERE MATRICE A DIAGONALISER (ORIGINE AU CENTRE
C***** DE GRAVITE DES VARIABLES)
      DO 5 J=1,NSB
C***** DISTANCE AU CENTRE DE GRAVITE DE LA MODALITE J
      DISTO(J)=(SOM/(PL(J)*PC) -1)
      DO 5 I=1,NSB
      RAC=SQRT(PL(I)*PL(J))
    5 X(I,J)=X(I,J)/(RAC*PC)-(RAC/SOM)
      PC=PC/SOM
      DO 751 I=1,NSB
      PL(I)=PL(I)/SOM
  751 CONTINUE
      NS=NSB
      WRITE(NSORT,666)
  666 FORMAT(/// 30H ANALYSE DES CORRESPONDANCES                    ///)
      NX=3
      NR=NSB-NVAR
      NCUM(1)=0
      DO 30 KK=2,NVAR
      NCUM(KK)=NCUM(KK-1)+NMOD(KK-1)
   30 CONTINUE
C***** CHOIX D UNE BASE DU SUPPORT DU NUAGE
       L=1
      K=2
      DO 22 J=1,NSB
      IF(J.EQ.NCUM(K))GO TO 21
      DO 20 I=1,NSB
   20 A(I,L)=X(I,J)
      L=L+1
      GO TO 22
   21 K=K+1
   22 CONTINUE
      DO 40 I=1,NSB
   40 WRITE(NSORT,41)(A(I,J),J=1,NR)
C***** ORTHONORMALISATION DE CETTE BASE
      CALL HILB(A,NR,NSB,NV)
C***** EXPRESSION DE X DANS CETTE NOUVELLE BASE
   41 FORMAT(1H 10F12.6)
      DO 24 I=1,NSB
      DO 24 J=1,NR
      VEC(I,J)=0
```

```
      DO 24 K=1,NSB
   24 VEC(I,J)=VEC(I,J)+X(I,K)*A(K,J)
      DO 25 I=1,NR
      DO 25 J=1,NR
      X(I,J)=0
      DO 25 K=1,NSB
   25 X(I,J)=X(I,J)+A(K,I)*VEC(K,J)
      CALL EIGNVA(NR,X,EIG,VEC,IND,NV)
      DO 27 J=1,NX
C***** EXPRESSION DES FACTEURS TROUVES DANS LA BASE INITIALE
      DO 27 I=1,NSB
      X(I,J)=0
      DO 26 K=1,NR
   26 X(I,J)=X(I,J)+A(I,K)*VEC(K,J)
   27 X(I,J)=(X(I,J)/SQRT(PL(I)))*SQRT(EIG(J))
      DO 43 I=1,NR
   43 WRITE(NSORT,41)(X(I,J),J=1,NR)
      WRITE(NSORT,10)
      WRITE(NSORT,9)(EIG(K),K=1,NR)
    9 FORMAT(1H 10F12.8)
   10 FORMAT(21H VALEURS PROPRES                    )
C  POURCENTAGES
  701 FORMAT(/  10H  TRACE=     F15.5)
      SOMV=0
      DO 700 J=1,NR
  700 SOMV=SOMV+EIG(J)
      WRITE(NSORT,701)SOMV
      DO 800 JM1=1,NR
      J=JM1
      POURC=(EIG(J)/SOMV)*100
  800 WRITE(NSORT,900)JM1,POURC
  900 FORMAT(/ 13H VAL.PROPRE    I3,15H  POURCENTAGE    F6.2)
C***** LES TROIS PREMIERS FACTEURS SONT F1,F2,F3.L ENSEMBLE DES FACTEURS
C***** FIGURENT DANS LES PREMIERES COLONNES DE X
      DO 199 I=1,NSB
      F1(I)=X(I,1)
      F2(I)=X(I,2)
      F3(I)=X(I,3)
  199 CONTINUE
C***** CALCUL DES AIDES A L INTERPRETATION,EDITIONS DES RESULTATS.
      CALL ICONT(X,EIG,JCAR,PL,NSB,DISTO,NV)
      RETURN
      END
```

```
      SUBROUTINE HILB(A,LCV,LCI,NV)
C***** ORTHONORMALISATION DES LCV PREMIERES COLONNES DE A
      DIMENSION A(NV,NV),VX(150)
      EPS=0.0000001
      DO 1 LV=1,LCV
      LV1=LV-1
      DO 2 LU=1,LV
      VX(LU)=0
      DO 3 I=1,LCI
    3 VX(LU)=VX(LU)+A(I,LV)*A(I,LU)
    2 CONTINUE
      IF(LV1.EQ.0)GO TO 4
      DO 5 I=1,LCI
      DO 5 LU=1,LV1
    5 A(I,LV)=A(I,LV)-VX(LU)*A(I,LU)
    4 CONTINUE
      TR=0
      DO 6 I=1,LCI
    6 TR=TR+A(I,LV)*A(I,LV)
      IF(TR/VX(LV)-EPS)8,8,22
    8 TR=0
      GO TO 23
   22 CONTINUE
      TR=1./SQRT(TR)
   23 CONTINUE
      DO 7 I=1,LCI
    7 A(I,LV)=A(I,LV)*TR
    1 CONTINUE
      RETURN
      END


      SUBROUTINE ICONT(VEC,EIG,JCAR,PL,NSB,DISTO,NV)
C***** CALCUL ET EDITION DES CONTRIBUTIONS ABSOLUES,RELATIVES,DES
C***** MASSES RELATIVES DES MODALITES,DES COORDONNEES SUR LES AXES,
C***** DES DISTANCES AU CENTRE DE GRAVITE
      DIMENSION EIG(1),JCAR(1),PL(1),DISTO(1)
      DIMENSION VEC(NV,NV)
      DIMENSION CONTI(5),CONTRI(5)
      COMMON/ENSOR/NINT,NSORT
      WRITE(NSORT,464)
      WRITE(NSORT,465)
  464 FORMAT(//,1X,4HNOMS,3X,6HMASSES,5X,5HDISTO,18X,11HCOORDONNEES,21X,
     122HCONTRIBUTIONS ABSOLUES,10X,23HCONTRIBUTIONS RELATIVES      )
  465 FORMAT(// 35X,2HF1,10X,2HF2,10X,2HF3,13X,2HF1,7X,2HF2,7X,2HF3,
     110X,2HF1,7X,2HF2,7X,2HF3,     /)
      NX=3
      DO 462 I=1,NSB
      DO 460 NFAC=1,NX
      CONTI(NFAC)=(VEC(I,NFAC  )**2)*PL(I)
      CONTRI(NFAC)=CONTI(NFAC)/(DISTO(I)*PL(I))
      CONTI(NFAC)=(CONTI(NFAC)/EIG(NFAC  ))*100.
  460 CONTINUE
      III=I
  462 WRITE(NSORT,463)JCAR(III),PL(I),DISTO(I),(VEC(I,K),K=1,3),
     1(CONTI(NY),NY=1,NX),(CONTRI(NZ),NZ=1,NX)
  463 FORMAT(1X,A3,3X,F6.3,2X,F10.4,2X,3F12.4,4X,3F9.2,4X,3F9.3)
      RETURN
      END
```

```
      SUBROUTINE GUSB(Y,VEC,EIG,G,NVAR,NIND,NX,NV,NMOD,NTOU)
C***** CALCUL DES COORDONNEES DES INDIVIDUS SUR LES NX PREMIERS AXES
C***** FACTORIELS
      DIMENSION Y(1),G(NX,NIND)
      DIMENSION VEC(NV,NV),EIG(1),NMOD(1)
      COMMON/BBB/NB
C***** LE TABLEAU EIG NE CONTI ENT QUE LES VALEURS PROPRES NON TRIVIALES
C***** NIND EST LE NOMBRE D INDIVIDUS,NVAR LE NOMBRE DE VARIABLES
C*****  LES COORDONNEES CALCULEES ONT ICI POUR VARIANCE 1
      DO 1 J=1,NIND
      READ(NB)(Y(K),K=1,NTOU)
      DO 1 NY=1,NX
      NY1=NY
      COEFF=1.0/(NVAR*EIG(NY1))
      ADEB=0.000001
      GG=0
      DO 2 L=1,NVAR
      NU=Y(L)+ADEB
      GG=GG+VEC(NU,NY1)
    2 ADEB=ADEB+NMOD(L)
    1 G(NY,J)=GG*COEFF
      RETURN
      END



      SUBROUTINE PLUSB(Y,VEC,G,NIND,NVAR,NMOD,PL,NX,NSB,NV,NTOU)
C*****  ICI,NVAR EST LE NOMBRE DE VARIABLES SUPPLEMENTAIRES
C***** NSB EST LE NOMBRE DE MODALITES SUPPLEMENTAIRES
C*****  NTOU EST LE NOMBRE TOTAL DE VARIABLES
C*****  N4 EST LE NOMBRE DE VARIABLES DEJA ANALYSEES
C*****  LE TABLEAU G EST ISSU DU SOUS PROGRAMME  GUSB
      DIMENSION Y(1),G(NX,NIND)
      DIMENSION VEC(NV,NX)
      DIMENSION NMOD(1),PL(1)
      COMMON/BBB/NB
      N4=NTOU-NVAR
      DO 4 I=1,NSB
      PL(I)=0
      DO 4 J=1,NX
    4 VEC(I,J)=0
      DO 2 J=1,NIND
      ADEB=0.00000001
      READ(NB)(Y(L),L=1,NTOU)
      DO 1 L=1,NVAR
      NU=ADEB+Y(L+N4)
      PL(NU)=PL(NU)+1
      DO 5 NY=1,NX
      VEC(NU,NY)=VEC(NU,NY)+G(NY,J)
    5 CONTINUE
    1 ADEB=ADEB+NMOD(L+N4)
    2 CONTINUE
      DO 8 I=1,NSB
      DO 8 J=1,NX
    8 VEC(I,J)=VEC(I,J)/PL(I)
      RETURN
      END
```

```
      SUBROUTINE PERMU(T1,T2,NB1,NB2,JC1,IMAX,NTOU,NM1,NOUV)
C***** CE SOUS PROGRAMME TRANSFERT DE LA BANDE NB1 A LA BANDE NB2
C***** UNE PERMUTATION DES VARIABLES SOUS CODAGE REDUIT,COMMANDEE PAR LE
C***** TABLEAU NOUV(K)
C***** IL CALCULE EGALEMENT LES NOUVEAUX IDENTIFICATEURS DES MODALITES
C*****  A L ENTREE JC1 EST LE TABLEAU DES IDENTIFICATEURS DES MODALITES
      DIMENSION JC1(1)
C***** A L ENTREE,NM1(K) EST LE NOMBRE DE MODALITE DE L ANCIENNE VAR.K
      DIMENSION T1(1),T2(1),NM1(1),NOUV(1)
      DIMENSION NM2(100),JC2(300)
      DIMENSION NCUM(100)
C    ATTENTION      AUCUN  REWIND  DANS  CE    SOUS-PROGRAMME
C NOUV(K)=L   SIGNIFIE QUE LA VALEUR SITUEE EN L DANS T1 VA EN K DANS T2
C***** SI NB2 = 0 , LA COPIE DU FICHIER N A PAS LIEU
      IF(NB2.EQ.0) GO TO 100
      DO 1 I=1,IMAX
      READ(NB1)(T1(J),J=1,NTOU)
      DO 2 J=1,NTOU
      L=NOUV(J)
    2 T2(J)=T1(L)
      WRITE(NB2)(T2(J),J=1,NTOU)
    1 CONTINUE
  100 CONTINUE
C***** LA RELATIVE COMPLEXITE DANS LA PERMUTATION DES IDENTIFICATEURS
C***** EST DUE AU NOMBRE QUELCONQUE DE MODALITES DES VARIABLES
      LDEB=0
      DO 3 J=1,NTOU
      NCUM(J)=0
      L=NOUV(J)
      NM2(J)=NM1(L)
      NCUM(J)=LDEB
    3 LDEB=LDEB+NM1(J)
      LDEB=0
      DO 4 J=1,NTOU
      NMM=NM2(J)
      DO 5 K=1,NMM
      L=NOUV(J)
      LL=NCUM(L)
    5 JC2(LDEB+K)=JC1(LL+K)
    4 LDEB=LDEB+NM2(J)
      DO 6 I=1,LDEB
    6 JC1(I)=JC2(I)
      DO 7 J=1,NTOU
    7 NM1(J)=NM2(J)
C****A LA SORTIE ,JC1 ET NM1 CORRESPONDENT AU NOUVEL ORDRE DES VARIABLES
      RETURN
      END
```

```
      SUBROUTINE IMPI(A,JCAR,NSB,NV)
C***** IMPRIME LE TABLEAU DE BURT A,AVEC NSB MODALITES,IDENTIFIEEES
C***** PAR JCAR..NV MAJORE NSB.
      DIMENSION A(NV,NV),JCAR(1)
      COMMON/ENSOR/NINT,NSORT
      NS=30
      DO 450 K=1,NSB,NS
       NL=K+NS-1
      IF(NSB.LT.NL)NL=NSB
      WRITE(NSORT,456)(JCAR(I),I=K,NL)
  456 FORMAT(1H /1H  4X,30(1X,A3))
      DO 451 I=1,NSB
      NF=NL
      IF(NF.GE.K)WRITE(NSORT,455)JCAR(I),(A(I,J),J=K,NF)
  455 FORMAT(1H /1H A3,1H 30F4.0)
  451 CONTINUE
  450 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE PARTI(K,IMAX,JMAX,NMOD,NITER,IDENT,T,LX,KLAS,JCAR,ICAR)
      DIMENSION NCENT(50),PJI(150),PJ(150),POID(150)
      DIMENSION TAB(150,16)
      DIMENSION NMOD(1),IDENT(1),T(1),LX(1),KLAS(1)
      DIMENSION JCAR(1)
      DIMENSION ICAR(1)
      COMMON/AGRA/TAB
      COMMON/ENSOR/NINT,NSORT
      COMMON/TOUTOU/NTOU
      COMMON/BBB/NB,NBB
C*****  K EST LE NOMBRE MAXIMUM DE CLASSES DEMANDEES
C***** IMAX = NOMBRE D INDIVIDUS
C***** JMAX = NOMBRE TOTAL DE MODALITES
C*****  NMOD(K)=NOMBRE DE MODALITES DE LA VARIABLE K
C***** NITER= NOMBRE MAXIMUM D ITERATION PREVU
C***** IDENT=IDENTIFICATEURS DES INDIVIDUS
C***** JCAR=IDENTIFICATEURS DES MODALITES
C***** ICAR = IDENTIFICATEURS (CONVENTIONNELS) DES CLASSES
C***** LES GERMES(CENTRES PROVISOIRES DE CLASSES) SONT DANS  TAB
C***** A LA SORTIE,TAB EST UN TABLEAU DE CONTINGENCE (K,JMAX)
      DO 14 J=1,JMAX
   14 PJ(J)=0
      EMAX=IMAX
      PI=1./EMAX
C***** TIRAGE (EXHAUSTIF) DES PREMIERS CENTRES DE CLASSE
      CALL TIRA(K,IMAX,NCENT,LX)
      WRITE(NSORT,101)(NCENT(M),M=1,K)
  101 FORMAT(20H INDIVIDUS CHOISIS =    6I10)
  121 FORMAT(11F7.3)
      NA=IMAX
      L=1
      SOM=0
      REWIND NB
C*****  PREMIER REMPLISSAGE DE TAB AVEC LES INDIVIDUS TIRES AU HASARD
      DO 1 I=1,IMAX
      READ(NB)(T(J),J=1,NTOU)
C***** CODAGE BINAIRE DES MODALITES
      CALL ETAL(T,NMOD,PJI,JMAX,NTOU)
      DO 10 J=1,JMAX
      PJ(J)=PJ(J)+PJI(J)
   10 CONTINUE
      SOM=IMAX*JMAX
      DO 2 M=1,K
      IF(I-NCENT(M))2,3,2
    2 CONTINUE
      GO TO 1
    3 CONTINUE
      DO 4 J=1,JMAX
    4 TAB(J,L)=PJI(J)
      WRITE(NSORT,103)I,L
      WRITE(NSORT,102)(TAB(J,L),J=1,JMAX)
  103 FORMAT(5H  I=  I5,5H L=      I2,15H   TAB(J,L)              )
  102 FORMAT(1X,30F4.0)
      POID(L)=PI
      L=L+1
    1 CONTINUE
C***** BALAYAGES SUCCESSIFS DU TABLEAU
      DO 12 NIT=1,NITER
```

```
      REWIND NB
      CALL GROUP(PJI,PI,PJ,TAB,KLAS,IMAX,JMAX,POID,SOM,K,T,NMOD)
      WRITE(NSORT,13)NIT
   13 FORMAT(///10X,20H ITERATION NUMERO                    I3,///)
      CALL JMPR(KLAS,IMAX,IDENT,K)
      WRITE(NSORT,100)(ICAR(KL),KL=1,K)
  100 FORMAT(/// 4X,30(1X,A3))
      DO 15 J=1,JMAX
   15 WRITE(NSORT,16)JCAR(J),(TAB(J,L),L=1,K)
   16 FORMAT(1H A3,30F4.0)
   12 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE GROUP(PJI,PI,PJ,TAB,KLAS,IMAX,JMAX,POID,SOM,K,T,NMOD)
      DIMENSION PJI(1),PJ(1),POID(1),KLAS(1),NMOD(1)
      DIMENSION T(1)
      DIMENSION PP(150)
      DIMENSION TAB(150,16),TAB2(150,16),POID2(150)
      COMMON/ENSOR/NINT,NSORT
      COMMON/BBB/NB,NBB
      COMMON/TOUTOU/NTOU
C*****    PROTECTION
      DO 12 J=1,JMAX
      IF(PJ(J))17,17,12
   17 WRITE(NSORT,19)J
   19 FORMAT(//22H ATTENTION LA VARIABLE  I5,15H A UN POIDS NUL  //)
      GO TO 100
   12 PP(J)=SQRT(SOM/PJ(J))
      DO 10 L=1,K
      DO 11 J=1,JMAX
      TAB(J,L)=(TAB(J,L)/POID(L))*PP(J)
   11 TAB2(J,L)=0
   10 POID2(L)=0
      API=1./PI
      DO 1 I=1,IMAX
C***** LECTURES ,POUR L INDIVIDU I ,DES VARIABLES J
      READ(NB)(T(J),J=1,NTOU)
C***** CODAGE BINAIRE DE CES VARIABLES
      CALL ETAL(T,NMOD,PJI,JMAX,NTOU)
      DO 15 J=1,JMAX
   15 PJI(J)=PJI(J)*API*PP(J)
      L=1
C***** CALCULS DES DISTANCES ENTRE LES INDIVIDUS ET LES CENTRES
      CALL        DIST(TAB,PJI,JMAX,L,DISTA)
      A=DISTA
      MIN=1
      DO 2 L=2,K
      CALL        DIST(TAB,PJI,JMAX,L,DISTA)
      B=DISTA
      IF(A-B)2,2,4
    4 MIN=L
      A=B
    2 CONTINUE
      KLAS(I)=MIN
C***** AFFECTATION DE I AU CENTRE LE PLUS PROCHE
      DO 5 J=1,JMAX
      PJI(J)=PJI(J)/(API*PP(J)     )
    5 TAB2(J,MIN)=TAB2(J,MIN)+PJI(J)
      POID2(MIN)=POID2(MIN)+PI
    1 CONTINUE
      DO 13 L=1,K
      DO 14 J=1,JMAX
   14 TAB(J,L)=TAB2(J,L)
   13 POID(L)=POID2(L)
  100 RETURN
      END
```

```
      SUBROUTINE DIST(TAB,PJI,JMAX,L,DISTA)
      DIMENSION TAB(150,16),PJI(1)
C***** CALCUL DE LA DISTANCE ( DU KHI-2) DE L INDIVIDU I AU GERME L
      DISTA=0
      DO 1 J=1,JMAX
      A= PJI(J)-TAB(J,L)
    1 DISTA=DISTA+A*A
      RETURN
       END
      SUBROUTINE ETAL(Y,NMOD,X,JMAX,NTOU)
C***** CODAGE BINAIRE DU VECTEUR DE DESCRIPTION
      DIMENSION Y(1),NMOD(1),X(1)
      DO 1 I=1,JMAX
    1 X(I)=0
      ADEB=0.000001
      DO 2 J=1,NTOU
      L=ADEB+Y(J)
      X(L)=1.
    2 ADEB=ADEB+NMOD(J)
      RETURN
      END
```

```
      SUBROUTINE JMPR(KLAS,IMAX,IDENT,K)
C***** EDITIONS DES RESULTATS.IDENTIFICATIONS DES INDIVIDUS
C***** APPARTENANT AU DIFFERENTES CLASSES
      DIMENSION KLAS(1),NUM(1514),IDENT(1)
      COMMON/ENSOR/NINT,NSORT
      DO 1 L=1,K
      LZ=1
      LA=1
      NEFF=0
      DO 2 I=1,IMAX
      IF(KLAS(I)-L)2,3,2
    3 NUM(LZ)=IDENT(I)
      LZ=LZ+1
      NEFF=NEFF+1
    2 CONTINUE
      WRITE(NSORT,4)L,NEFF
    4 FORMAT(/16H CLASSE NUMERO      ,I3,20H    EFFECTIF =          ,I4)
      LZ=LZ-1
      IF(LZ)1,1,6
    6 CONTINUE
      WRITE(NSORT,5)(NUM(II),II=LA,LZ)
    5 FORMAT(20(2H *,I4))
    1 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE TIRA(K,IMAX,NCENT,LX)
      DIMENSION LX(1)
      DIMENSION NCENT(1)
      COMMON/ENSOR/NINT,NSORT
C***** TIRAGE ALEATOIRE (SUR CDC 6600) DES K PREMIERS GERMES
      DO 2 I=1,IMAX
    2 LX(I)=I
      DO 1 L=1,K
      IML=IMAX-L+1
      I=IML*RANF(0.0)+1
      NCENT(L)=LX(I)
      DO 3 M=I,IML
      LX(M)=LX(M+1)
    3 CONTINUE
    1 CONTINUE
      WRITE(NSORT,10)(NCENT(L),L=1,K)
   10 FORMAT(/// 1H 12I10,/)
      RETURN
      END
```

VI - EXEMPLE D'APPLICATION

Vue d'ensemble des caractéristiques des familles constituant l'échantillon de l'enquête CNAF-CREDOC 1971.

Ce paragraphe résumera les caractéristiques démographiques et socio-économiques de l'ensemble des ménages en tenant compte des corrélations existant entre elles. Nous nous limiterons aux variables que l'on peut qualifier de "variables les plus immédiates", c'est-à-dire aux variables ayant servi à la construction même de l'échantillon, également à celles qui permettent de mieux résumer la position et l'insertion sociale d'une famille.

Les âges, les professions du père et de la mère, leurs salaires respectifs, le nombre d'enfants et l'âge de l'aîné, la catégorie de commune de résidence, l'appartenance à certaines associations seront donc simultanément prises en compte pour tenter de dégager une espèce de "trame" de l'échantillon, sur laquelle pourront être tissés les différents thèmes constituant le contenu même de l'enquête : attitudes et comportements des ménages vis-à-vis de la familles, des équipements sociaux, etc...

Nous tenterons de synthétiser de façon parlante toutes ces informations dans l'étude qui va suivre, dont l'essentiel sera résumé par le graphique ci-après.

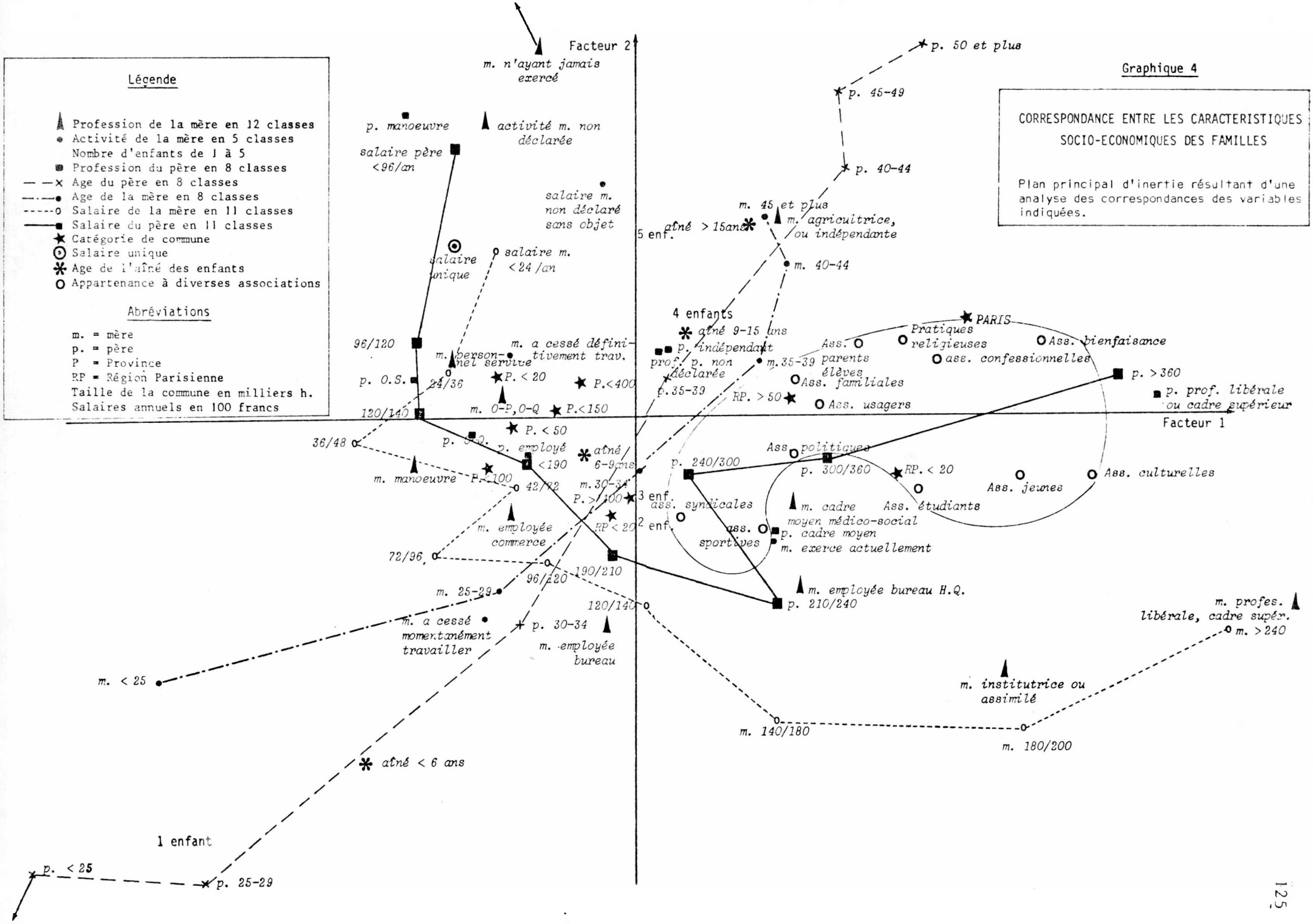
Disons que ce graphique a l'ambition, non pas de contenir autant d'informations que tous les tableaux croisés réalisables à partir des variables étudiées, mais de faire ressortir les grandes lignes du réseau d'interrelations existant entre ces variables.

Nous verrons quels enseignements nous apporte cette représentation, qui nous renverra alors, pour certains approfondissements ou précisions, vers des tableaux qui n'auraient pas été consultés, ou qui n'auraient pas retenu notre attention.

Graphique 4

CORRESPONDANCE ENTRE LES CARACTERISTIQUES SOCIO-ECONOMIQUES DES FAMILLES

Plan principal d'inertie résultant d'une analyse des correspondances des variables indiquées.



Codage et traitement de l'information

24 variables toutes qualitatives (1) ont fait l'objet du traitement statistique. Ces variables présentant diverses modalités (nombre d'enfants : 5 modalités). En tout il y a 118 modalités possibles de réponse pour un ménage.

Ainsi, chacun des 1545 ménages étudiés ici sera caractérisé par une suite de 118 "oui" ou "non" selon qu'il satisfait ou non à telle ou telle modalité de réponse. Il y a un seul "oui" pour une variable (question) donnée ; par exemple il y a 8 modalités de réponse pour la profession du père, il y aura obligatoirement un "oui" et sept "non" (on ne peut appartenir qu'à une seule catégorie socio-professionnelle). Chaque modalité sera également caractérisée par une suite de 1545 "oui" et "non", indiquant les diverses réponses des ménages.

Pour le traitement numérique de l'information, les "oui" seront codés : 1 et les non : 0. On conçoit, à partir de ce codage, qu'il soit possible de définir des distances entre les diverses catégories de réponses : deux modalités seront très voisines si les ménages qui ont répondu "oui" à l'une ont également répondu "oui" à l'autre. Inversement, elles seront éloignées si chaque ménage répond "oui" à l'une et "non" à l'autre.

Ce sont ces distances que la représentation ci-dessous essaie de traduire concrètement : cependant, ce graphique ne sera qu'une approximation, car les distances statistiques n'ont pas les propriétés géométriques requises pour donner lieu à une figure plane.

---

(1) L'analyse porte sur l'échantillon non redressé.

8 variables sont des variables de structure pour l'analyse, ce sont :

1 : la profession du père	5 : l'âge de l'aîné
2 : l'activité de la mère	6 : l'âge de la mère
3 : la profession de la mère	7 : le salaire du père
4 : le nombre d'enfants	8 : le salaire de la mère

16 servent à illustrer la typologie précédente :
1 : âge du père - 2 : catégorie de commune - 3 : différence d'âge homme-femme - 4 : perception du salaire unique - 5 : pratique religieuse - 6 à 16 : appartenance à des associations : familiales, parents d'élèves, syndicales, bienfaisance, politiques, confessionnelles, étudiants, culturelles, jeunes, sportives, usagers.

Notons que certaines des variables analysées ont des modalités ordonnées de façon naturelle (par exemple, la variable "âge de la mère" comprend 8 modalités ; si l'on excepte la modalité "non-réponse, sans objet", il existe un ordre naturel des différentes classes d'âge). Sur le graphique, ces modalités ordonnées sont jointes par un trait continu, afin de permettre de suivre facilement l'évolution du phénomène continu sous-jacent.

Interprétation générale

Caractéristiques classiques

Cet éparpillement de variables, apparemment assez confus, s'ordonne en réalité autour de deux grands thèmes : l'âge de la famille et son statut social.

Suivons en effet les diverses classes d'âge du père, depuis le bas gauche du graphique jusqu'au haut droit, suivant une diagonale assez rectiligne : le long de cet axe, on trouve également les classes d'âge croissantes de la mère, moins dispersées toutefois que celles de leur mari ; on trouve également le nombre d'enfants croissant progressivement, avec un léger infléchissement sur la gauche, correspondant au fait que les familles les plus nombreuses se trouvent surtout dans les milieux modestes ; on trouve également, le long de ce même axe, les différentes classes d'âge de l'aîné des enfants, également rangées par ordre croissant dans la même direction.

Assez perpendiculairement à cette direction se trouvent les lignes brisées joignant les différentes classes de salaire du père et de celui de la mère lorsqu'elle travaille. Ces deux lignes brisées sont reliées à leurs extrémités, et tournent leur concavité vers le haut. Les extrémités correspondent aux statuts sociaux précisément les plus extrêmes, ce repliement vers le haut nous montre que ce sont les familles plutôt âgées qui occupent les situations sociales les plus divergentes : aisance ou dénuement.

Le long de ces lignes brisées, les catégories socio-professionnelles du père et de la mère viennent illustrer et confirmer ce trajet le long de l'échelle sociale : aux manoeuvres, gens de maison et ouvriers de diverses catégories à gauche s'opposent sur la droite les professions libé-

rales et les cadres supérieurs.

On peut noter le décalage extrêmement important existant entre les classes de salaire des femmes et celles des hommes situées à proximité sur le graphique. On peut également noter un net écartement des courbes pour les hauts salaires, la courbe des salaires des pères semblant attirée vers le haut : la "baïonnette" que l'on observe sur la ligne brisée "salaire du père" semble correspondre au phénomène suivant : 53% des femmes dont le mari a un salaire annuel compris entre 21000 et 24000 francs exercent effectivement une profession, alors que ce pourcentage tombe à 33% pour les femmes dont le mari a un salaire annuel compris entre 24000 et 33000 francs (rappelons que nous étudions ici l'échantillon non redressé). Le point représentant cette classe de familles a donc tendance à se rapprocher du point "salaire unique", situé dans le quadrant haut et gauche du graphique.

Les positions des points représentant les diverses catégories de communes nous montrent que Paris et la région parisienne sont plutôt situés sur la droite du graphique, alors que la province, particulièrement les petites villes, occupe une région du graphique moins favorisée.

Cette situation est à rapprocher du tableau suivant :

SALAIRE DU PERE SELON L'ORIGINE GEOGRAPHIQUE

Salaire du père (annuel)	Province (pour 1000)	Région parisienne (pour 1000)
Moins de 9600 Francs	63	36
De 9600 à 12000 Francs	133	36
De 12000 à 14400 Francs	165	114
De 14400 à 16800 Francs	151	91
De 16800 à 19200 Francs	99	94
De 19200 à 21600 Francs	64	96
De 21600 à 24000 Francs	43	68
De 24000 à 30000 Francs	62	104
De 30000 à 36000 Francs	41	68
Plus de 36000 Francs	69	190

Les bas salaires sont donc beaucoup plus fréquents en province, alors que les hauts salaires du père se trouvent surtout dans la région parisienne.

Participation et adhésion à des associations

L'ensemble des points représentant l'adhésion à des associations a été entouré d'une ligne continue afin de faire ressortir leur relative concentration.

Il s'en faut beaucoup que la répartition soit uniforme sur l'ensemble de l'échantillon. La lecture de ce graphique nous montre que les adhérents de ces associations pourtant très diverses sont surtout des familles d'un statut social plutôt élevé. Cependant, à l'exception des associations de parents d'élèves et des associations syndicales, elles ne mettent en jeu que des effectifs trop faibles d'adhérents pour que l'on puisse donner des estimations fiables et procéder à une analyse fine.

On peut cependant, en regroupant les salaires du père en deux classes, arriver à mettre en évidence directement ce phénomène.

APPARTENANCE A DES ASSOCIATIONS SELON LE SALAIRE DU PERE

	Salaire annuel du père < 30000 F.		Salaire annuel du père > 30000 F.	
Effectifs totaux	1 156		227	
Adhérents à :				
Association Familiale	63	5%	23	10%
Association parents d'élèves	372	32%	128	56%
Association syndicale	330	28%	51	22%
Association bienfaisance	37	3%	25	11%
Association politique	43	4%	13	6%
Association confessionnelle	59	5%	36	16%

Si l'on excepte les associations syndicales, on voit que le pourcentage d'adhérents est toujours plus élevé pour les classes de revenus supérieurs. Il est même trois plus fort pour les associations de bienfaisance et les associations confessionnelles.

Signalons la position du point "salaire unique" qui devrait occuper une position plus centrale dans le nuage de points, vu l'équilibre des ménages vis-à-vis de ce critère. En fait on a vu que le taux d'activité des mères de famille est plus élevé aux statuts socio-professionnels supérieurs. Les femmes les plus qualifiées ont à la fois plus d'avantages à exercer leur activité et peuvent plus facilement subvenir aux frais de garde des enfants, grâce à leur rémunération. En outre la distorsion entre la perception du salaire unique et l'activité de la mère est plus importante chez les mères les moins qualifiées. C'est surtout le cas chez les mères dont le salaire mensuel est inférieur à 300 francs. Il s'agit sans doute d'activités d'appoint très aléatoires.

CHAPITRE III

DESCRIPTION STATISTIQUE DE CERTAINES RELATIONS BINAIRES

(Analyse des correspondances locales)

GENERALITES

La méthode d'analyse des correspondances, considérée comme un algorithme de réduction de données, permet d'obtenir des représentations visuelles à partir de tableaux de valeurs numériques positives et nous fournit des règles d'interprétation permettant d'inférer la structure ou certains traits pertinents des données de base à partir de ces représentations.

De façon un peu simplificatrice, disons que la technique opère comme un appareil radiographique (l'opacité des tissus, obstacle à la vision directe du squelette et des organes étant alors l'analogue du caractère multidimensionnel des données, obstacle à leur assimilation).

L'utilisation d'un tel appareil nécessite une certaine formation technique de l'utilisateur, éventuellement une certaine préparation des matériaux à radiographier, enfin, une interprétation des clichés obtenus liée au principe même de fonctionnement de l'appareil, et également à l'entraînement et à l'expérience du radiologue. Il ne suffit pas de savoir comment l'opacité aux rayons X dépend de la densité, du volume, de la composition chimique des organes pour pouvoir identifier ou même simplement voir une tâche pulmonaire.

L'analyse des correspondances exige également une expérience clinique de l'utilisateur. Celui-ci doit notamment savoir ce qu'il faut attendre de la méthode lorsque le tableau d'entrée a une structure relativement fruste, lorsque certaines situations typiques se présentent.

En ce sens, l'étude de relations binaires particulières permet de procéder à un étalonnage de la méthode de lecture des résultats des analyses.

Nous verrons ainsi dans une première partie quelles représentations sont obtenues lorsque le tableau analysé est la matrice associée à certains graphes particuliers. Malgré leur caractère artificiel, nous verrons que ces procédures d'étalonnage nous permettront de faire quelques remarques utiles. Elles permettront de préciser certaines des notions introduites dans la seconde partie où nous nous préoccupons de relations binaires existant "a priori" entre des observations statistiques.

Dans cette seconde partie, on généralisera également certains traitements qui évaluent la compatibilité existant entre une partition "a priori" d'un ensemble I et une correspondance sur I x J, J désignant un ensemble de descripteurs. Le cas d'une relation binaire symétrique non transitive entre éléments de I sera étudié.

La première partie fait largement appel à des résultats contenus dans la note de J.-P. BENZECRI citée en référence (1).

I - DESCRIPTION DE CERTAINS GRAPHES A PARTIR DE LEURS MATRICES CARACTERISTIQUES

I-1. Rappels et notations.

Nous commencerons par rappeler quelques définitions en précisant nos notations.

- Un graphe $G = (X,U)$ est défini par la donnée d'un ensemble X, dont les éléments s'appellent les sommets, $X = (x_1, x_2, \ldots x_n)$ et d'une famille U d'éléments du produit cartésien XxX ; $U = (u_1, u_2, \ldots u_p)$ Les élements de U sont les arcs de G.

Nous nous intéresserons surtout aux graphes symétriques ou non-orientés.

Pour ces graphes, si $u = (x,y) \in U$, alors $(y,x) \in U$.

L'ensemble à deux éléments (x,y) (qui n'est pas un sous ensemble de X, puisque x peut être égal à y) sera noté e_i, et constituera la i-ème arête du graphe $G = (X,E)$, avec $E = (e_1, e_2, \ldots e_m)$.

Dans le cas où deux sommets quelconques de G ne sont reliés que par au plus une arête, la famille E est une partie symétrique de XxX. On appelle matrice associée au graphe G non-orienté la matrice carrée M d'ordre (n,n), de terme général m_{ij}, avec $m_{ij} = k$ si les sommets x_i et x_j sont reliés par k arêtes (ou k-adjacents, ou k-contigüs) $m_{ij} = 0$ sinon.

On appelle matrice d'incidence du graphe G la matrice T d'ordre (n,m) de terme général t_{ik}, avec $t_{ik} = 1$ si le sommet x_i est adjacent à l'arête e_k, $t_{ik} = 0$ sinon.

Enfin on appelle degré $d(x_i)$ d'un sommet x_i de G le nombre d'arêtes ayant au moins une extrémité en x_i.

On désignera par N la matrice diagonale d'ordre (n,n) dont le i-ème élément diagonal vaut $d(x_i)$. N est la matrice des degrés. Le graphe est dit "homogène de degré r". Si $d(x_i) = r$ pour tout i, on a alors $N = rI_n$. (I_n étant la matrice unité d'ordre n).

Les graphes que nous rencontrerons dans les applications seront des graphes simples, c'est-à-dire des graphes non-orientés, sans boucles (arêtes du type (x,x)), avec, pour tout couple de sommets (x_i, x_j), $m_{ij} \leq 1$.

(Exemple : graphe des communes de la région parisienne, où deux sommets sont reliés par une arête si les communes correspondantes sont contiguës).

Nous serons cependant conduits à étudier des graphes plus généraux construits à partir de ceux-ci.

I-2. Propriétés des matrices M,T,N, relatives à un graphe sans boucle G=(X,E).

a) Les matrices (N + M) et (N - M) sont définies non-négatives

En effet, il résulte de la définition de la matrice d'incidence que, en désignant par $\tilde{T}$ la transposée de T : $T\tilde{T} = N+M$, ce qui établit la première partie de l'assertion.

La seconde partie découle du fait que la forme quadratique $\tilde{X}(N-M)X$ est une variance (cf. §I-3).

b) Si φ est un facteur de norme 1 issu de l'analyse des correspondances du tableau M, relatif à la valeur propre $\lambda(\varphi) \neq 0$, et si $\varepsilon(\varphi)$ vaut +1 ou -1 selon la parité de φ, alors φ est également facteur de norme 1 issu de l'analyse des correspondances de T, relativement à la valeur propre $\dfrac{1 + \varepsilon(\varphi)\sqrt{\lambda(\varphi)}}{2}$.

En effet, l'analyse de M nous conduit à la relation de transition

$$N^{-1}M\varphi = \varepsilon(\varphi)\sqrt{\lambda(\varphi)} .$$

Celle de T nous conduit à l'équation :

$$\frac{1}{2} N^{-1}T\tilde{T}\Psi = \mu\Psi \qquad \text{avec} \quad T\tilde{T} = N+M$$

On en déduit : $N^{-1}M\psi = (2\mu - 1)\psi$

d'où $\mu = \dfrac{1 + 2(\varphi)\sqrt{\lambda\ (\varphi)}}{2}$

c) Si le graphe G est bichromatique (ou encore biparti), chaque valeur propre $\lambda(\varphi)$ issue de l'analyse de M correspond à au moins un facteur direct et un facteur inverse.

On peut en effet grouper les sommets de même couleur de façon à diviser M en quatre blocs, les deux blocs diagonaux étant formés uniquement de 0. Il est clair, sous cette forme, que $N^{-1}M$ admet des valeurs propres deux à deux opposées. On passe d'un facteur direct φ au facteur inverse correspondant à la même valeur propre en changeant les signes des composantes φ_i de φ pour une seule couleur.

Remarque : La matrice d'incidence d'un tel graphe G n'est autre que le tableau des réponses à un questionnaire formé de deux questions mises sous forme disjonctive complète. Le tableau N+M est le tableau de contingence de BURT correspondant. L'analyse se réduit de toute façon à la diagonalisation d'une matrice d'ordre au plus égal au plus petit nombre de sommets de la même couleur.

Rappelons qu'un graphe est bichromatique si et seulement si il n'admet pas de cycle de longueur impaire.

Ainsi les arbres, les réseaux à mailles carrées ou hexagonales, etc...doivent être décrits par l'analyse des tableaux T ou N+M ou par diagonalisation directe de $N^{-1}M$. En effet, l'analyse de M nous donnerait, pour chaque $\lambda(\varphi)$, une base d'un sous espace propre engendré par deux facteurs dont l'un seulement est direct : les vecteurs propres extraits de cette façon risquent de n'avoir pas grand sens.

[Pour un questionnaire comprenant p questions mises sous forme disjonctive complète , le tableau des réponses peut être considéré comme la matrice d'incidence d'un hypergraphe uniforme de rang p, admettant p pour nombre chromatique fort.

Une arête d'un hypergraphe H(X,E) est une partie non vide de X, dont le cardinal peut être supérieur à 2. L'hypergraphe est dit uniforme de rang p si toutes les arêtes ont pour cardinal p.

(Si p=2, on retrouve la notion du graphe simple). Enfin, le nombre chromatique fort est une des généralisations du nombre chromatique des graphes qui implique que toutes les couleurs des sommets d'une même arête soient différentes].

I-3. Variance locale d'une fonction sur X. Optimalité des facteurs φ

Soit G = (X,E) un graphe non-orienté sans boucle à n sommets, dont les matrices caractéristiques sont encore notées M,T,N.

Soit Y une application de X dans $\mathbb{R}$, et y le point de R^n, dont les composantes sont les images par Y des sommets de X.

Nous désignerons par m la somme des éléments de M (ou la trace de N): le nombre m vaut deux fois le nombre d'arêtes (cardinal E) de G. C'est le nombre d'arcs de G considéré comme un graphe symétrique.

On appelle variance locale de y la quantité :

$$v_\ell(y) = \frac{1}{2m} \sum_{(i,j) \in E} (y_i - y_j)^2$$

Il s'agit donc du demi écart quadratique moyen calculé sur les m accroissements (chaque arête intervenant deux fois dans la sommation) correspondant à des sommets adjacents sur le graphe.

La variance empirique des composantes de y vaut, si $\bar{y}$ désigne la moyenne arithmétique des y_i :

$$v_t(y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{2n(n-1)} \sum_{i,j} (y_i - y_j)^2$$

- GEARY a défini (réf. 5,6) le coefficient de contiguïté de la fonction y comme le quotient $C(y) = v_\ell(y) / v_t(y)$

  Ce quotient est voisin de 1 si les valeurs des y_i ne dépendent pas de G. Il est inférieur à 1 s'il existe une certaine continuité de leur répartition sur G. Pour le calcul des caractéristiques de la loi de C(y) (dans l'hypothèse d'indépendance vis-à-vis du graphe) en fonction de sa matrice associée M, on pourra consulter la référence 6.

En notation matricielle, on a :

$$v_\ell(y) = \frac{1}{m}\cdot\tilde{y}(N-M)y$$

Si U désigne la matrice d'ordre (n,n) associée à la n-clique = ($u_{ij} = 1$ pour tout i et $j \leqslant n$)

$$v_t(y) = \frac{1}{n(n-1)}\,\tilde{y}(nI-U)y$$

- Nous modifierons légèrement la définition de $v_t(y)$, en affectant à chaque valeur y_i un poids p_i proportionnel au degré n_i du sommet i : $p_i = n_i/m$

Dans ces conditions : $\bar{y} = \frac{1}{m}\sum n_i y_i$ et $v_t(y) = \frac{1}{m}\sum_{i=1}^{n} n_i(y_i-\bar{y})^2$

Les deux définitions coïncident si le graphe est homogène.

Le coefficient de contiguïté étant une quantité homogène de degré 0 en y, et invariant par translation parallèlement à la droite des fonctions constantes (première bissectrice)

$$y\in\mathbb{R}^n,\quad a \neq 0,\quad a \text{ et } b\in\mathbb{R} \Longrightarrow C(ay + b) = C(y),$$

on pourra se limiter à l'étude des fonctions y^* de moyennes nulles et de variance 1.

Propriété 1 : la fonction y^* dont le coefficient de contiguïté est minimal est le facteur direct φ de l'analyse des correspondances du tableau M correspondant à la plus grande valeur propre $\lambda(\varphi)$, si celle-ci est unique.

On a alors : $C(y^*) = 1 - \sqrt{\lambda(\varphi)}$

Si la plus grande valeur propre λ correspondant à un facteur direct n'est pas unique, toute fonction du sous-espace propre correspondant est une fonction dont la contiguïté est minimale.

Plus généralement, le sous-espace à r dimensions tel que toute fonction $y^*\in R$ soit de contiguïté minimale est constitué par les r facteurs directs issus de l'ana-

lyse de M correspondant aux r plus grandes valeurs propres associées à des facteurs directs.

On a en effet :

$$C(y^*) = \frac{\tilde{y}^* (N-M) y^*}{\tilde{y}^* N y^*}$$

Chercher le minimum de C revient à chercher la plus petite valeur de ν pour laquelle $(N-M)y^* = \nu\, Ny^*$

ou encore : $N^{-1}My^* = (1-\nu)y^*$

Chercher le minimum de ν revient à chercher le maximum de $1-\nu$, qui vaut $\varepsilon(\varphi)\sqrt{\lambda(\varphi)}$

On a dans tous les cas : $C(\varphi) = 1 - \varepsilon(\varphi)\sqrt{\lambda(\varphi)} \quad \varphi \in \Phi$
Ce qui nous donne une caractérisation des facteurs directs et inverses:

Propriété 2 : Un facteur est direct (resp : inverse) si sa variance locale est inférieure (resp : supérieure) à sa variance totale.

La propriété 1 de contiguïté minimale des premiers facteurs nous garantit une reconstitution géométrique satisfaisante des graphes qui correspondent à des structures géométriques simples, dont nous verrons quelques exemples plus loin.

De plus, dans la généralisation de l'analyse discriminante proposée dans la seconde partie, ces facteurs feront figure de cas-limite. De la même façon, une variable indicatrice d'une partition (facteur φ relatif à la valeur propre 1 de la matrice associée au graphe décrivant la partition par des cliques disjointes, avec $C(\varphi) = 0$) est la meilleure fonction discriminante possible.

La distinction entre facteurs directs et inverses peut paraître artificielle (cf. réf. 3) puisque, comme nous l'avons vu, l'analyse de

N+M redonne les mêmes facteurs que celle de M, tous directs, classés selon la croissance de leur coefficient de contiguïté. Cependant, le fait d'ajouter des boucles en chaque sommet "étouffe" la structure géométrique du graphe. Rappelons que M^r est la matrice associée au multigraphe G^r ayant autant d'arêtes joignant x_i à x_j qu'il y a de chemins de longueur r dans G joignant x_i à x_j.

L'existence de facteurs inverses correspondant à des valeurs propres élevées est une conséquence des conflits pouvant exister entre les distances 1 et 2 dans le graphe G, c'est-à-dire entre (i et j sont adjacents) et (i et j ont les mêmes adjacents).

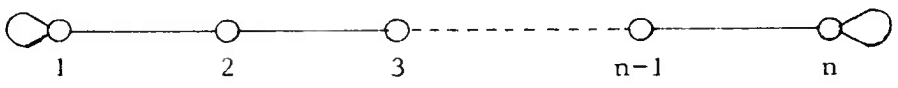
Ce conflit entre M et M^2 est exhibé par le produit terme à terme de ces deux matrices, qui n'est autre que trM^3, c'est-à-dire trois fois le nombre de triangles de G. D'ailleurs $(trM^3)^2 / (trM^2)^3$ caractérise l'assymétrie (coefficient de K. PEARSON) de la répartition des valeurs propres de M autour de l'origine.

I-4. Description de graphes particuliers (cf. réf. 1)

Nous nous limiterons à trois types de graphes homogènes (de degré constant) :

- chaîne ayant des boucles aux extrémités, à n sommets
- cycle de longueur n .
- réseau à mailles carrées, qui peut être défini comme la somme cartésienne de deux chaînes à n_1 et n_2 sommets.

I-4.1 - Description d'une chaîne homogène :



La relation $N^{-1}M\varphi = \varepsilon(\varphi)\sqrt{\lambda(\varphi)}\,\varphi$ demande la résolution de l'équation aux différences finies

$$\frac{1}{2}(\varphi_{j-1} + \varphi_{j+1}) = \varepsilon(\varphi)\sqrt{\lambda(\varphi)}\,\varphi_j$$

dont l'équation caractéristique est, en posant $\epsilon(\varphi)\sqrt{\lambda(\varphi)} = \lambda'$

$$\mu^2 - 2\lambda'\mu + 1 = 0$$

Comme $\lambda' < 1$, cette équation n'a que des racines complexes, de module 1.

Une base des solutions est donc fournie par les fonctions du type : $\alpha \cos j\omega + \beta \sin j\omega$, avec $\lambda' = \cos\omega$.

Les conditions aux limites nous donnent

$$\omega = \frac{2p\pi}{2n+1} \qquad \alpha = \frac{\cos p\pi}{2n+1} \qquad \beta = \frac{-\sin p\pi}{2n+1}$$

Ainsi pour toutes les valeurs de $p \leq n$, la suite de fonction sur X :

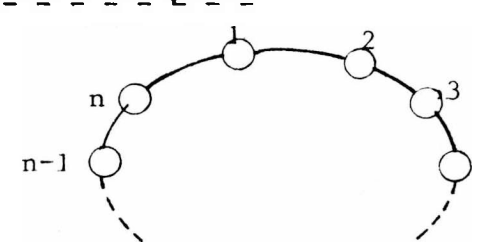
$$\varphi^p(j) = \cos\left(\frac{(2j+1)p\pi}{2n+1}\right)$$

est une suite de facteurs relatifs respectivement aux valeurs propres :

$$\lambda_p(\varphi) = \cos^2\left(\frac{2p\pi}{2n+1}\right)$$

Le premier facteur restitue l'ordre des sommets sur le graphe, ce qui était prévisible à partir des résultats précédents puisque toute interversion accroîtrait la variance locale sans changer la variance totale. Les facteurs suivants sont des fonctions polynomiales du premier.

I-4.2 - Description d'un cycle.



Seules, les conditions aux limites changent par rapport aux calculs précédents :

les fonctions $\varphi^p(j) = \cos\left(\frac{2jp\pi}{n}\right)$ et $\psi^p(j) = \sin\left(\frac{2jp\pi}{n}\right)$

sont des facteurs relatifs à la valeur propre $\cos^2\left(\frac{2p\pi}{n}\right)$.

Dans le plan des deux premiers facteurs, on obtient l'équation paramétrique d'un cercle.

I-4.3. - Description d'un réseau à mailles carrées

a) Somme cartésienne de deux graphes :

Soit $G = (X,E)$ et $H = (Y,F)$ deux graphes non-orientés ayant respectivement n_G et n_H sommets de matrices associées notées également G et H.

On suppose que $g_{ij} \leqslant 1$ et $h_{kl} \leqslant 1$

(G et H ne sont pas des graphes simples, car chaque sommet peut avoir une boucle).

On appelle somme cartésienne le graphe G + H dont l'ensemble des sommets est le produit cartésien des ensembles X et Y, et où deux sommets (x,y) et $(x',y') \in XxY$ sont adjacents si et seulement si l'une ou l'autre des deux conditions suivantes est réalisée.

(1) $x = x'$ et $(y,y') \in F$

(2) $y = y'$ et $(x,x') \in E$

De plus, si ces deux conditions sont réalisées simultanément (cas où les sommets x et y ont chacun une boucle) alors le sommet (x,y) de G+H a deux boucles.

Notons $A \otimes B$ le produit tensoriel des matrices A et B :

$$(a \otimes b)i,k;j,l = a_{ij}b_{kl}$$

et désignons par S la matrice associée au graphe G + H :

On a immédiatement : $S = I_{n_G} \otimes H + G \otimes I_{n_H}$

$$s(i,k;j,l) = \delta_i^j \cdot h_{kl} + g_{ij} \cdot \delta_k^l$$

Supposons les graphes G et H homogènes de degré r_G et r_H

Si φ est un facteur issu de l'analyse de G, relatif à la valeur propre $\lambda(\varphi)$, ψ un facteur issu de l'analyse de H, relatif à la valeur propre $\mu(\psi)$

On a : $$\frac{1}{r_G} G\varphi = \varepsilon(\varphi)\sqrt{\lambda(\varphi)} \cdot \varphi$$

et $$\frac{1}{r_H} H\psi = \varepsilon(\psi)\sqrt{\mu(\psi)} \cdot \psi$$

Soit θ une fonction sur XxY telle que : $\theta_{jk} = \varphi_i \psi_k$

θ vérifie l'équation : $\frac{1}{r+g} S\theta = \alpha\theta$

avec : $$\alpha = \frac{r_G \varepsilon(\varphi)\sqrt{\lambda(\varphi)} + r_H (\psi)\sqrt{\mu(\psi)}}{r_G + r_H}$$

On a bien : $$\sum_{jl} s(i,k;j,l)\theta_{jl} = \varphi_i \sum_l h_{kl}\psi_l + \psi_k \sum_j g_{ij}\varphi_j$$

Or $$\sum_l h_{kl}\psi_l = r_H \varepsilon(\psi)\sqrt{\mu(\psi)}\,\psi_k$$

et $$\sum_j g_{ij}\varphi_j = r_G \cdot \varepsilon(\varphi)\sqrt{\lambda(\varphi)}\,\varphi_i$$

On a donc :

$$\sum_{jl} s(i,k;j,l)\theta_{jl} = \theta_{ik}\,(r_G \varepsilon(\varphi)\sqrt{\lambda(\varphi)} + r_H \varepsilon(\psi)\sqrt{\mu(\psi)})$$

Ainsi, les facteurs et les valeurs propres de l'analyse des correspondances d'une somme cartésienne de deux graphes S = G + H se déduisent aisément des facteurs et des valeurs propres des analyses des correspondances de G et de H lorsque les graphes G et H sont homogènes.

b) Analyse d'un réseau rectangulaire à mailles carrées homogène

Un tel réseau (dont les sommets figurant sur les côtés et les coins sont munis de boucles de façon à assurer un degré constant) peut être considéré comme une somme cartésienne de graphes $S = G + H$ où G et H sont des chaînes homogènes (ayant respectivement n_1 et n_2 sommets) telles que celles qui ont été étudiées en 1.

Le paragraphe précédent nous montre que :

$$\theta^{(p,q)}_{ik} = \cos \frac{(2i+1)p\pi}{2n_1+1} \cos \frac{(2k+1)q\pi}{2n_2+1}$$

est un facteur de l'analyse de S, relatif à la valeur propre

$$\lambda(p,q) = \frac{1}{4}\left[\cos \frac{2p\pi}{2n_1+1} + \cos \frac{2q\pi}{2n_2+1}\right]^2$$

I-5. Remarques générales

Les descriptions précédentes, jointes à des résultats empiriques (concernant par exemple la description du graphe des départements français, où les positions respectives des départements sont correctement restituées dans le plan des deux premiers facteurs) nous prouvent que lorsque la structure de graphe est compatible avec une représentation euclidienne de l'ensemble des sommets (i.e. si l'adjacence représente bien un certain type de voisinage définissable à partir d'une métrique euclidienne), les premiers facteurs directs de l'analyse des correspondances de la matrice associée au graphe restituent convenablement cette représentation.

Des voies de recherche restent ouvertes pour préciser des mots tels que "voisinage" et "convenablement"...

En fait, il s'agit seulement de restituer des positions relatives plus que des distances. Cependant, l'économie du codage reste considérable : ainsi, dans le cas des départements français, la comparaison avec l'analyse de proximité est intéressante. A la donnée d'une ordonnance, c'est-à-dire d'une suite de (90 x 89/2) = 4005 couples de sommets ordonnés

suivant leurs distances croissantes, on substitue un ensemble de 422 couples non ordonnés (l'homogénéité du "semis" des départements nous conduit à penser que ces 422 couples font partie des 500 premiers éléments de l'ordonnance) qui sont les arêtes de notre graphe.

On peut noter le fait que les valeurs propres ont une décroissance relativement lente (cf. par exemple, dans le cas du cycle, la décroissance de $\cos^2(2p\pi/n)$, pour des valeurs élevées de n, alors que les deux premiers facteurs restituent de façon satisfaisante la configuration initiale).

On aurait pu s'attendre à voir les deux premières valeurs propres prédominer, puisque l'essentiel de la structure est décrit par une figure plane.

En fait, avec ce type de codage, l'inertie extraite par un facteur, dont on sait qu'elle donne en général une idée pessimiste de l'information résumée par ce facteur, est pratiquement sans rapport avec son pouvoir explicatif.

Désignons par M la matrice associée au graphe G du cycle dont chaque sommet est muni d'une boucle - Pour $\alpha \in {}^{+}$ la matrice M^{α}, table de contingence symétrique dont l'élément $m_{ij}^{(\alpha)}$ représente le nombre de chemins de longueur α joignant le sommet i au sommet j dans G, décrit de façon beaucoup plus globale le graphe initial - Les deux premiers facteurs issus de l'analyse de M^{α} sont évidemment les mêmes que ceux de M mais le taux d'explication des deux premières valeurs propres peut être rendu voisin de 1 lorsque α est grand.

Un élément d'appréciation intéressant est également fourni par l'examen de la façon dont la suite des vecteurs propres dépend des premiers. Nous avons vu que dans le cas d'une chaîne, ou d'un cycle, ils sont des fonctions polynomiales des premiers. Ce type de variation est généralement observé lors de la description de graphes empiriques. Dans le cas où le nombre de sommets du graphe augmente indéfiniment, il a été mis en évidence (cf. réf. 1) que l'analyse du graphe rejoint la recherche des fonctions propres de l'opérateur de LAPLACE. Les formes analytiques de ces fonctions relativement à certains domaines géométriques simples sont répertoriées, ainsi que les valeurs propres correspondantes, et la concordance avec les résultats d'analyses empiriques s'avère satisfaisante. Dans le cas de structures représentables dans un espace à p dimensions, on désigne

par "effet GUTMANN à p dimensions" la dépendance analytique des facteurs d'ordre supérieur à p vis à vis des premiers facteurs.

II - RELATIONS BINAIRES "A PRIORI" SUR UN ENSEMBLE D'OBSERVATIONS STATISTIQUES

Il est fréquent que des observations multidimensionnelles se répartissent en classes disjointes, cette partition étant connue de façon purement exogène. En analyse des données, on travaille généralement sur un corpus de mesures (dont on sait qu'il doit être homogène et dans la mesure du possible exhaustif vis-à-vis du phénomène étudié) et l'on illustre les représentations obtenues en traçant les centres de gravités des mesures appartenant à une même classe. Ainsi, on étudie par exemple une typologie des ménages à partir de leurs profils de consommation, et l'on fait apparaître la partition de ces ménages en catégories socio-professionnelles en projetant par exemple dans le plan des premiers facteurs d'une analyse les centres de gravité et les caractéristiques de dispersion de chacune des classes. Dans un second temps, on peut chercher à reconstituer au mieux les classes à partir des profils, c'est-à-dire procéder à une analyse discriminante.

Il arrive cependant que l'on ne puisse répartir les observations en classes disjointes, autrement dit que notre connaissance exogène du phénomène se traduise par une relation binaire qui ne soit pas forcément transitive. Ainsi, une enquête de consommation est faite sur un échantillon de ménages répartis dans 80 agglomérations, il est assez arbitraire de regrouper ces agglomérations par régions en vue de faire apparaître un éventuel effet géographique ; on est en effet conduit à ne prendre qu'un très petit nombre de régions (pour ne pas avoir de sous-échantillons trop restreints). Il est alors fréquent que deux agglomérations appartenant à des régions différentes soient plus proches géographiquement que deux agglomérations appartenant à une même région. Il paraît plus pertinent de retenir une relation de voisinage entre certaines agglomérations, qui peut d'ailleurs tenir compte du relief, des communications, etc...

Il en est de même lors du traitement des données se rapportant à des communes, des cantons ou des départements, pour lesquelles les analyses de données fournissent un bilan global. Si l'on n'utilise pas l'information que constituent les positions respectives des zones de prélèvements statistiques les unes par rapport aux autres, on ne mettra pas en évidence les distorsions ou éventuellement les conflits existant entre les liaisons au niveau local et les liaisons globales.

L'analyse discriminante se généralise aisément au cas des relations binaires symétriques quelconques. Nous étudierons ensuite le cas d'une correspondance entre deux ensembles dont l'un est muni a priori d'une structure de graphe symétrique.

II-1. Variables aléatoires sur les sommets d'un graphe symétrique

Le modèle statistique le plus simple susceptible de générer des observations dépendant de la structure de graphe stipule l'existence d'un coefficient d'autocorrélation entre deux observations adjacentes, les observations non adjacentes étant non corrélées. Explicitons ce modèle.

Si $G = (X,E)$ désigne un graphe simple à n sommets de matrice associée M, un vecteur aléatoire Y, dont les n composantes correspondant chacune à un sommet de G ont une espérance mathématique commune (μ) sera défini comme "dépendant du graphe G" si sa matrice des covariances théoriques s'écrit :

$$V = \sigma^2 (I + \rho M)$$

Si le graphe G est homogène de degré r, la condition pour que V soit définie non négative est que $|\rho| \leqslant \frac{1}{r}$, puisque les valeurs propres de M sont inférieures à r en valeur absolue.

Bien entendu, cette condition peut être améliorée pour certains types de graphes.

Il est fréquent que des dépendances de ce type soient suggérées par le champ des observations : c'est le cas des observations chronologiques (G est une chaîne), des observations correspondant à diverses classes (G est formé de cliques disjointes), des observations géographiques (G est planaire).

Il est alors clair que l'effectif n des observations n'est pas comparable à la taille d'un échantillon formé d'observations indépendantes. Nous définirons une "pseudo-taille" de l'ensemble des observations dépendant du graphe G comme la taille de l'échantillon d'observations indépendantes qui donnerait la même information sur la moyenne théorique μ commune à toutes les observations.

Désignons par Y le vecteur aléatoire se réalisant sur le graphe $G = (X,U)$; nous supposerons que Y est un vecteur normal de matrice des covariances $V = \sigma^2(I + \rho M)$ avec $E(y_i) = \mu$ pour tout $i \leq n$. E désigne l'espérance mathématique, et $|V|$ le déterminant de V. La densité de probabilité du vecteur Y s'écrit alors, dans le cas où V est régulière :

$$P(Y) = \frac{1}{(2\pi)^{\frac{n}{2}} |V|^{\frac{1}{2}}} \exp. \left\{ -\frac{1}{2}(\widetilde{Y-\mu})\, V^{-1}(Y-\mu) \right\}$$

On a : $\mathcal{L} = \text{Log } P(Y) = -\frac{n}{2}\text{Log } 2\pi + \frac{1}{2}\text{ Log } \left|V^{-1}\right| - \frac{1}{2}\text{ tr}\left[V^{-1}(X-\mu)(\widetilde{X-\mu}) \right]$

La quantité d'information de FISHER relative à la moyenne μ s'écrit

$$I_\mu = E\left(\frac{\partial^2 \mathcal{L}}{\partial \mu^2} \right) = \text{tr} V^{-1} U$$ (où U désigne la matrice associée à la n-clique : $u_{ij} = 1$ pour tout i et pour tout j)

En exprimant V à partir de la matrice associée au graphe :

$$I = \frac{1}{\sigma^2}\ \text{tr}(I + M)^{-1}U = \frac{1}{\sigma^2}\ \text{tr}(I - \rho M + \rho^2 M^2 - \rho^3 M^3 + \ldots.)U$$

Or $\text{tr}M^k U$ n'est autre que la somme des éléments de M^k, matrice associée au multigraphe des distances k, c'est donc deux fois le nombre total C_k de chemins de longueur k du graphe G.

Donc : $$I = \frac{1}{\sigma^2}(n - 2C_1\rho + 2C_2\rho^2 - 2C_3\rho^3 + \ldots..)$$

Pour $\rho = 0$, on retrouve la quantité connue $I_\mu = \frac{n}{\sigma^2}$.

On vérifie que si ρ est négatif, la quantité d'information sur la moyenne (les autres paramètres étant supposés connus) est plus élevée que si les observations sont indépendantes. Elle est évidemment plus faible

si $\rho > 0$. Nous appellerons "pseudo-taille" de l'échantillon l'effectif fictif n_p de réalisation indépendante donnant la même quantité d'information sur la moyenne.

$$n_p = tr(I + M)^{-1}U = n - 2\sum_k C_k \rho^k$$

Vérifions la cohérence de cette définition pour différents types de graphes simples :

1) Graphes formés de doublets disjoints :

Le nombre de chemins distincts de longueur k est égal à $\frac{n}{2}$ quel que soit k

d'où : $$n_p = n(1 - \rho + \rho^2 .....) = \frac{n}{1+\rho}$$

Si ρ tend vers 1, alors $n_p \longrightarrow \frac{n}{2}$.

Du point de vue de l'information sur la moyenne, il y a alors deux fois moins d'observations.

2) Graphe complet sans boucle :

Le nombre de chemins de longueur k est : $\frac{n(n-1)^k}{2}$

d'où : $$n_p = n(1 - (n-1)\rho + (n-1)^2\rho^2 - .....)$$

$$n_p = \frac{n}{1+(n-1)\rho}$$

On peut dans ce cas calculer directement $(I + \rho U)^{-1}$ ($u_{ij} = 1$ pour tout i et j).
Si $\rho \to 1$, le nombre fictif d'observations tend aussi vers 1.

Remarquons que la matrice $(I + \rho U)$ a pour plus grande valeur propre $1+(n-1)\rho$, toutes les autres étant égales à $1 - \rho$; cette matrice est définie positive quel que soit ρ tel que : $-\frac{1}{n-1} < \rho < 1$. Il est donc loisible de supposer que ρ peut être aussi voisin de 1 que l'on veut.

3) Graphe formé de q composantes connexes complètes sans boucle (la i-ème composante ayant n_i sommets).

On a immédiatement, V^{-1} s'inversant par blocs ;

$$n_p = \sum_{i=1}^{q} \frac{n_i}{1+(n_i-1)\rho} \qquad (n_p \to q \text{ quand } \rho \to 1)$$

Cette formule contient les deux précédentes.

II-2. Etude de l'inertie locale

II-2.1 - Définitions :

Nous étudierons le cas d'un vecteur de description à p composantes caractérisant chacun des n sommets d'un graphe $G = (X,E)$, de matrice associée M et de matrice des degrés N.

Y désignera maintenant une application de X dans R^p, $p \in N^+$. y_i^j est la valeur du descripteur j (ou de la variable j) pour le sommet i.

L'homologue de la variance locale definie précédemment est la matrice des covariances locales V_ℓ, de terme général, pour deux descripteurs j et j' :

$$v_\ell(j,j') = \frac{1}{2m} \sum_{(i,i')\in E} \left(y_i^j - y_{i'}^j\right)\left(y_i^{j'} - y_{i'}^{j'}\right)$$

(m désigne toujours deux fois le nombre d'arêtes de G).

Si Y désigne également le tableau à n lignes et p colonnes des valeurs de l'application, on a :

$$V_\ell = \frac{1}{m} \tilde{Y}(N-M)Y$$

La matrice des covariances empiriques s'écrit de la même façon, la sommation étant cette fois effectuée sur tous les couples (i, i') de sommets.

$$v_t(j,j') = \frac{1}{2n(n-1)} \sum_{i,i'} (y_i^j - y_{i'}^j)(y_i^{j'} - y_{i'}^{j'})$$

Soit u une application linéaire de R^p dans R -

On note u(i) l'image d'un point i de R^p par u.

$$u(i) = \sum_j u_j y_i^j$$

Par analogie avec la terminologie économique, on appellera indice descriptif du sommet x_i l'image u(i) de l'application composée de X dans R : uoY.

La variance locale d'un indice descriptif u est donnée par la forme quadratique :

$$v_\ell(u) = \frac{1}{2m} \sum_{(i,i')\in E} \left[u(i)-u(i')\right]^2$$

Soit, en notation matricielle :

$$v_\ell(u) = \tilde{u} V_\ell u$$

Le coefficient de contiguïté d'un indice descriptif u s'écrit alors comme le quotient :

$$C(u) = \frac{\tilde{u} V_\ell u}{\tilde{u} V_t u}$$

II-2.2 - Cas d'une correspondance.

Soit f_{ij} une correspondance sur IxJ, produit cartésien de deux ensembles finis (de cardinaux encore notés I et J), avec $\sum_{i\in I} \sum_{j\in J} f_{ij} = 1$

L'ensemble I est supposé coïncider ou être en bijection avec l'ensemble X des sommets d'un graphe G = (X,E).
On pose $f_{i.} = \sum_j f_{ij}$ et $f_{.j} = \sum_i f_{ij}$

A chaque sommet $i \in I$ correspond un point-profil de R^j dont les coordonnées sont $(f_{ij}/f_{i.})_{j\in J}$ et une masse $f_{i.}$.

On appelle de même indice descriptif d'un sommet i la valeur $u(i)$

$$u(i) = \sum_{j \in J} u_j \frac{p_{ij}}{p_{i.}}$$

La variance totale d'un indice s'écrit alors :

$$v_t(u) = \frac{1}{2} \sum_{i \in I} \sum_{i' \in I} f_{i.} f_{i'.} (u(i)-u(i'))^2$$

soit :

$$v_t(u) = \frac{1}{2} \sum_{i \in I} \sum_{i' \in I} f_{i.} f_{i'.} \left[\sum_{j \in J} u_j \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right) \right]^2$$

$$v_t(u) = \frac{1}{2} \sum_{j \in J} \sum_{j' \in J} u_j u_{j'} \left[\sum_{i \in I} \sum_{i' \in I} f_{i.} f_{i'.} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right) \left(\frac{f_{ij'}}{f_{i.}} - \frac{f_{i'j'}}{f_{i'.}} \right) \right]$$

$= \tilde{u} V_t u$. V_t, matrice d'inertie totale, a pour terme général $v_t(j,j')$,

avec $v_t(j,j') = \frac{1}{2} \sum_{i \in I} \sum_{i' \in I} f_{i.} f_{i'.} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right) \left(\frac{f_{ij'}}{f_{i.}} - \frac{f_{i'j'}}{f_{i'.}} \right)$

On a également $v_t(j,j') = \sum_{i \in I} f_{i.} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right) \left(\frac{f_{ij'}}{f_{i.}} - f_{.j'} \right)$

Comme précédemment, la variance locale s'obtient en extrayant de la double sommation les couples d'indices correspondant à des sommets adjacents sur le graphe : la matrice d'inertie locale V_ℓ a alors pour terme général :

$$v_\ell(j,j') = \frac{1}{2m_\ell} \sum_{(i,i') \in E} f_{i.} f_{i'.} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right) \left(\frac{f_{ij'}}{f_{i.}} - \frac{f_{i'j'}}{f_{i'.}} \right)$$

avec maintenant : $m_\ell = \sum_{(i,i') \in E} f_{i.} f_{i'.}$

Désignons par matrice d'inertie différée V_d la matrice de terme général :

$$v_d(j,j') = \frac{1}{2m_d} \sum_{(i,i') \notin E} f_{i.} f_{i'.} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right) \left(\frac{f_{ij'}}{f_{i.}} - \frac{f_{i'j'}}{f_{i'.}} \right)$$

avec $m_d = \sum_{(i,i') \notin E} f_{i.} f_{i'.}$

On a alors la relation :

$$V_t = m_\ell V_\ell + m_d V_d$$

Cette décomposition coïncide avec la relation de HUYGHENS lorsque le graphe est formé de cliques disjointes (graphe d'une partition).

V_ℓ est alors la matrice d'inertie intraclasse

V_d devient la matrice interclasse.

On a : $\tilde{u} V_t u = m_\ell \tilde{u} V_\ell u + m_d \tilde{u} V_d u$;

L'inertie totale d'un indice descriptif u se décompose en inertie locale et inertie différée.

II-3. Généralisation de l'analyse discriminante

L'étude de la décomposition simultanée des deux formes quadratiques $\tilde{u}V_d u$ et $\tilde{u}V_t u$ dans le cas où G est un graphe symétrique quelconque (qui consiste à chercher une forme u telle que le quotient $\frac{\tilde{u}V_d u}{\tilde{u}V_t u}$ soit maximal, ou telle que $C(u) = \frac{\tilde{u}V_\ell u}{\tilde{u}V_t u}$ soit minimal), généralise la recherche des fonctions discriminantes calculées lorsque G est le graphe d'une partition : Il s'agit maintenant de chercher les indices descriptifs u dont le coefficient de contiguïté est minimal, c'est-à-dire ayant la répartition la plus continue sur le graphe. On désignera ce type d'analyse par "analyse locale".

Soit G un graphe à n sommets homogènes de degré r, de matrice associée M. Chaque sommet de G est décrit par p variables, colonnes

du tableau Z d'ordre (n,p). Comme les variances locales et totales ne dépendent pas de leurs moyennes, on peut supposer que les colonnes de Z sont centrées. Afin d'alléger les notations, nous supposerons que tous les sommets de G ont le même poids.

Dans ces conditions, le coefficient de contiguïté de l'indice descriptif u, qu'il s'agit de rendre minimal, s'écrit :

$$C(u) = \frac{1}{r} \cdot \frac{\tilde{u}\tilde{Z}\ (rI - M)\ Zu}{\tilde{u}\ \tilde{Z}Z\ u}$$

Ce qui nous conduit à l'équation, avec μ minimum :

$$\frac{1}{r}\ \tilde{Z}\ (rI - M)\ Zu = \mu\ \tilde{Z}Zu$$

Nous supposerons que $\tilde{Z}Z$ est non-singulière. La démonstration s'étendra immédiatement au cas où $\tilde{Z}Z$ et $\tilde{Z}(rI-M)Z$ ont même support, en faisant choix d'une base dans le support commun.

On a la relation : $\frac{1}{r}.\tilde{Z}MZu = (1-\mu)\tilde{Z}Zu$

Soit : $\frac{1}{r}(\tilde{Z}Z)^{-1}\tilde{Z}MZu = (1-\mu)\ u$

En posant $v = Zu$ (les n composantes de v sont les valeurs de l'indice descriptif u pour les sommets du graphe G)

$$\frac{1}{r}\left[Z(\tilde{Z}Z)^{-1}\tilde{Z}\right]Mv = (1-\mu)\ v$$

Le tableau entre crochets n'est autre que l'opérateur-projection P sur la variété linéaire $\mathcal{V}(Z)$ engendrée dans R^n par les p colonnes de Z.

Sous cette forme, l'intérêt et les limites de l'analyse discriminante (présentée ici dans un cadre plus général) apparaissent clairement :

1) Si le nombre des variables augmente et se rapproche de n, l'opérateur P se rapproche de l'identité, et les valeurs de l'indice v tendent vers les composantes de φ, vecteur propre de $\frac{1}{r}M$ correspondant à la plus grande valeur propre $(1 - \mu_{min})$. Ainsi, si G est le graphe d'une partition en deux classes, on peut améliorer progres-

sivement la discrimination en ajoutant des variables prises dans une table de nombres au hasard, jusqu'à obtenir φ (correspondant à $\mu_{min} = 0$), variable indicatrice de la partition, qui est dans ce cas vecteur propre de $\frac{1}{r}M$ correspondant à la valeur propre 1.

Il sera donc nécessaire (cf. réf. 2) de ne retenir que les dimensions pertinentes de la variété $\mathcal{V}(Z)$ en effectuant une analyse préalable du nuage des p points variables dans R^n, afin d'en éliminer le bruit.

Si G est un graphe associé à une carte géographique, nous avons vu dans la première partie que les deux premiers facteurs φ^1 et φ^2 de $\frac{1}{r}M$ constituent des coordonnées géographiques des sommets de G. Nous retrouverons donc cette carte si $\mathcal{V}(Z)$ contient deux points dont les composantes sont des coordonnées des sommets de G.

L'analyse locale permet de déceler les conflits existant entre la structure de graphe et les caractéristiques de descriptions des sommets. Accompagnée d'une analyse séparée des matrices V_t et V_ℓ elle permet de mettre en évidence une éventuelle hétérogénéité géographique des liaisons existant entre les variables. Nous verrons, en même temps qu'un exemple d'application, quelles sont les règles d'interprétation de ce type de technique et les procédures de calcul à mettre en oeuvre.

II-4. Lien avec l'étude des corrélations partielles

Dans le cas où le graphe étudié représente un ensemble de points d'un espace euclidien dont les positions peuvent être repérées par un petit nombre de coordonnées, il est tentant de rapprocher la notion de covariance locale à celle de covariance partielle, les variables-coordonnées étant supposées fixées.

Désignons par T le tableau à n lignes et q colonnes situant par q coordonnées les positions des n sommets du graphe $G = (X,E)$ homogène de degré r, de matrice associée M.

Les considérations théoriques et les résultats empiriques de la première partie nous prouvent que les q premiers facteurs directs

issus de l'analyse de M (colonnes de Φ_q^+) constituent dans une certaine mesure des coordonnées des sommets, de sorte que la variété linéaire $\mathcal{V}(T)$ engendrée par T dans R^{n-1} coïncide approximativement avec celle $\mathcal{V}(\Phi_q^+)$ engendrée par ces q facteurs - (Nous nous plaçons dans R^{n-1} car toutes les variables étudiées ici ont une moyenne nulle).

L'analyse des corrélations partielles du tableau Z (décrivant par p variables les n sommets de G), à situation géographique constante, revient à projeter le nuage de p points $\mathcal{N}(Z)$ sur le sous-espace orthogonal dans R^{n-1} de $\mathcal{V}(T)$.

L'opérateur projection sur ce sous-espace est représenté par la matrice $(I_n - T(T\tilde{T})^{-1}\tilde{T})$, et la matrice des covariances partielles s'écrit :

$$v_p(Z) = \frac{1}{n}.\tilde{Z}\ (I - T(T\tilde{T})^{-1}\tilde{T})Z$$

Alors que la matrice des covariances locales s'écrit :

$$v\ (Z) = \frac{1}{n}.\tilde{Z}\ (I - \frac{1}{r}M)Z$$

En fait, si les q premières valeurs propres de M était prédominantes, et les valeurs propres correspondantes voisines de 1, la formule de reconstitution nous donnerait :

$$\frac{1}{r}M \approx \Phi_q\ \tilde{\Phi}_q \approx T(T\tilde{T})^{-1}\tilde{T}$$

Nous savons qu'en fait, ces valeurs propres ne sont pas prédominantes; on a, en désignant par s la dimension du sous-espace permettant une reconstitution satisfaisante de M :

$$\frac{1}{r}\ M \approx \Phi_s \Lambda_s \tilde{\Phi}_s$$

Λ_s désignant la matrice diagonale dont les éléments diagonaux sont les s plus grandes valeurs propres de M.

- A une con traction (due aux valeurs propres) près, $\frac{1}{r}M$ agit donc comme un opérateur-projection sur le sous-espace $\mathcal{S}(T)$, engendré par $\mathcal{V}(T)$ et s-q vecteur dont on sait qu'ils sont des "fonctions polynomiales"

des précédents.

La covariance locale agit donc approximativement comme une covariance partielle après régression polynomiale sur les coordonnées.

Si z_i est une observation de la variable z correspondant au sommet i du graphe, et si $\vec{t}_i$ désigne le vecteur des coordonnées de ce sommet dans un espace euclidien R^p, nous pouvons supposer, plus généralement, que :

$$z_i = f(\vec{t}_i)+e_i$$

f étant une fonction continue à dérivées continues qui décrit la façon dont la valeur d'une variable dépend de sa position sur le graphe, e_i désigne la partie non explicable par la position.

On a alors pour deux sommets i et j adjacents sur le graphe, donc pour un vecteur $\overrightarrow{t_i - t_j}$ dont la longueur est petite :

$$z_i - z_j = e_i - e_j + (\overrightarrow{t_i - t_j}).\frac{\overrightarrow{\partial f}}{\partial t} + \ldots.$$

La partie principale des accroissements pris en compte dans le calcul des variances et covariances locales est constituée par les accroissements des termes résiduels, ce qui confirme l'interprétation de ces covariances en termes de covariances partielles ne mettant en jeu que des hypothèses faibles sur la nature de la liaison avec les variables que l'on désire fixer.

II-5. Programme d'analyse des correspondances locales

Ce programme étudie la décomposition simultanée des formes quadratiques d'inertie locales et totales, en tronquant à différents niveaux le support de la matrice d'inertie totale, de façon à éliminer le maximum de bruit.

Différents paramètres destinés à faciliter l'interprétation des résultats sont calculés.

a) Sous-programme_principal_AFCOT.

Arguments d'entrée :

X : Tableau de correspondance, ayant NSB lignes et NA colonnes.
Les paramètres NV et NO ne servent qu'à dimensionner.

JCAR : Tableau des identificateurs (en A3) des colonnes, puis des lignes de X.

Ce sont les colonnes de X qui représentent les sommets du graphe.

La matrice associée n'est jamais introduite dans la mémoire centrale : elle est condensée dans le tableau KA(I,J), qui est l'indice du Ième sommet adjacent au sommet J. (J varie de 1 à NA et I et 1 à NL, degré maximum d'un sommet du graphe).

Les paramètres d'entrée sont donc : X, NA, NSB, NO, NV, JCAR, KA, NL.

De plus certains paramètres sont introduits à partir du programme principal par des "COMMON" : NINT, NSORT : numéros logiques d'entrée et de sortie ; NTEST, NPAS : NTEST est le nombre maximum de dimensions que l'on désire abandonner pour "résumer" le support de la matrice d'inertie totale, NPAS définit les sauts qui vont permettre d'accèder à cette réduction maximale : ainsi, si par exemple NSB = 30, NPAS = 6, NTEST = 25, l'étude de la décomposition est faite d'abord avec 29 dimensions (support commun des deux formes quadratiques), puis en ne retenant respectivement que les 23, 17, 11, 5 premiers axes principaux d'inertie du nuage associé au tableau de correspondance X.

Les matrices d'inerties totale A et différée B sont calculées par l'appel de AFT2. Il faut prendre garde aux interversions des tableaux A et B lors de cet appel, et dans la suite du programme.

Sorties :

Les impressions des caractéristiques de diagonalisation sont faites par l'intermédiaire du sous-programme PROPRE, qui édite à chaque étape les valeurs propres, les vecteurs propres et les pourcentages d'inertie expliquée.

Sont imprimés successivement :

- lors de l'appel de AFT2 : la matrice d'inertie locale, la matrice des corrélations locales qui s'en déduit, les coefficients de contiguïté de chacune des NSB variables.

- Lors des appels de PROPRE, les caractéristiques spectrales de la matrice d'inertie totale, puis de la matrice symétrisée qui sert d'intermédiaire de calcul.

Directement à partir de AFCOT :

- les coefficients de contiguïté des indices descriptifs, les composantes de ces indices, leurs valeurs pour les NA sommets du graphe, différents paramètres destinés à vérifier la cohérence des calculs (dont le calcul pourra être omis dans une version moins expérimentale du programme),

- Les produits scalaires de l'indice avec les vecteurs unitaires portés par les NSB axes initiaux (analogues des contributions absolues), et les corrélations entre les valeurs des indices pour les NA sommets et les valeurs des variables pour ces sommets (analogues des contributions relatives).

b) Sous-programme AFT2 (utilise IMPA* , sous-programme éditant les tableaux).

Les arguments d'entrée sont le tableau de données X et ses marges PC et PL, ses dimensions réelles NSB et NA, et les valeurs maximales de ces dimensions NV et NO, le tableau d'identificateurs JCAR, le tableau KA codant la matrice associée au graphe, et NL, degré maximal d'un sommet du graphe. A la sortie, A désigne la matrice d'inertie différée et B la matrice d'inertie totale. Ces deux noms sont intervertis par la séquence d'appel.

c) Sous-programme BSPAS.

Ce sous-programme calcule, à partir d'une matrice symétrique B d'ordre (NSB,NSB), et d'une matrice de passage S, d'ordre (NSB,NFAC) la transformée $\tilde{S}BS$ qui est mémorisée en B.

Les listages des sous-programmes IMPA, EIGNVA, GRAPH, qui sont des auxiliaires usuels, ne figurent pas ci-dessous).

d) Sous-programme PROPRE : auxiliaire d'édition

e) Sous-programme EIGNVA : diagonalise une matrice symétrique.

f) Sous-programme GRAPH : représentation graphique.

```
      SUBROUTINE AFCOT(X,VEC,EIG,F1,F2,F3,NA,NSB,NO,NV,JCAR,KA,NL)
C*****  NA=NOMBRE DE SOMMETS DU GRAPHE
C***** NSB=NOMBRE DE DESCRIPTEURS DES SOMMETS
C***** NO ET NV  SONT  SONT DES MAJORANTS DE NA  ET  NSB
      DIMENSION VEC(NV,NV)
      DIMENSION X(NV,NO)
C*****  X EST LE TABLEAU DE DONNEES
      DIMENSION JCAR(1),EIG(1),F1(1),F2(1),F3(1)
      DIMENSION DISTO(100),PC(100),PL(100)
      DIMENSION A(50,50)
      DIMENSION B(50,50),S(50,50),PROD(30,100)
      DIMENSION KA(9,100)
C***** KA =TABLEAU  CODANT LA MATRICE ASSOCIEE AU GRAPHE
      DIMENSION LX(100)
      DIMENSION BNORM(50)
      COMMON/NORM/DENO
      COMMON/ENSOR/NINT,NSORT
      COMMON/REDUC/NTEST,NPAS
      SOM=0
      WRITE(NSORT,666)
  666 FORMAT(/// 40H ANALYSE DES CORRESPONDANCES LOCALE          ///)
C***** CALCUL DES MARGES ET DES FREQUENCES
      DO 1 I=1,NA
      PC(I)=0.
      DO 1 J=1,NSB
      PC(I)=PC(I)+X(J,I)
    1 SOM=SOM+X(J,I)
      DO 2 J=1,NSB
      PL(J)=0.
      DO 2 I=1,NA
      PL(J)=PL(J)+X(J,I)
    2 X(J,I)=X(J,I)/SOM
      DO 3 I=1,NA
    3 PC(I)=PC(I)/SOM
      DO 4 J=1,NSB
    4 PL(J)=PL(J)/SOM
C*****APPEL DE AFT2=A EST LA MATRICE D INERTIE TOTALE
C***** B EST LA MATRICE D INERTIE DIFFEREE
      CALL        AFT2(X,B,A,PC,PL,NA,NSB,NO,NV,JCAR,KA,NL)
      NW=NV
C     MAXIMUM DE  U B U  AVEC  U A U  FIXE
C*****CALCUL DES FACTEURS DU NUAGE TOTAL
      CALL EIGNVA(NSB,A,EIG,VEC,IND,NW)
      DO 8 I=1,NSB
      EIG(I)=ABS(EIG(I))
      DO 8 J=1,NSB
    8 VEC(J,I)=(VEC(J,I)/SQRT(PL(J)))*SQRT(EIG(I))
      NVEC=5
      WRITE(NSORT,800)
  800 FORMAT(// 40H FACTEURS DU NUAGE TOTAL                       ///)
      CALL PROP(EIG,NSB,VEC,NV,NVEC)
      NSB1=NSB-1
C*****  S  EST L OPERATEUR PROJECTION SUR LE SUPPORT PRINCIPAL DE  A
      DO 120 I=1,NSB
      DO 120 J=1,NSB1
  120 S(I,J)=VEC(I,J)*SQRT(PL(I))/EIG(J)
C***** B EST SAUVEGARDE DANS  A
      DO 200 I=1,NSB
```

```
      DO 200 J=1,NSB
  200 A(I,J)=B(I,J)
C***** BOUCLES AVEC DIFFERENTS TRONQUAGES DU SUPPORT
      DO 222 KCY=1,NTEST,NPAS
      NFAC=NSB-KCY
C***** LA DIMENSION RETENUE EST  NFAC
      DO 201 I=1,NSB
      DO 201 J=1,NSB
  201 B(I,J)=A(I,J)
C*****CALCUL DE LA TRANSFORMEE DE B  APRES TRONQUAGE
      CALL BSPAS(B,S,VEC,NV,NSB,NFAC)
      CALL EIGNVA(NFAC,B,EIG,VEC,IND,NV)
      CALL PROP(EIG,NFAC,VEC,NV,NFAC)
C***** CALCUL DES COEFFICIENTS DE CONTIGUITE DES FACTEURS EXTRAITS
      DO 205 J=1,NFAC
      EIG(J)=EIG(J)+(1-EIG(J))/DENO
  205 WRITE(NSORT,206)J,EIG(J)
  206 FORMAT(1H // 5H FACT  I4,30H (VAR.LOCALE)/(VAR.TOTALE)=    F8.3)
C***** EXPRESSION DES FACTEURS DANS LA BASE INITIALE
      DO 123 J=1,NFAC
      DO 123 I=1,NSB
      B(I,J)=0
      DO 124 K=1,NFAC
  124 B(I,J)=B(I,J)+S(I,K)*VEC(K,J)
  123 B(I,J)=B(I,J)/SQRT(PL(I))
      DO 125 J=1,NFAC
  125 WRITE(NSORT,126)(B(I,J),I=1,NSB)
  126 FORMAT(2H *  10F12.4)
C***** VALEURS DES FACTEURS(INDICES DESCRIPTIFS)POUR LES SOMMETS
C***** DU GRAPHE
      WRITE(NSORT,15)
   15 FORMAT(1H / 23H VALEURS DES FACTEURS                    )
      NX=NFAC
      DO 301 I=1,NSB
      LX(I)=JCAR(I+NA)
  301 CONTINUE
      DO 13 J=1,NA
      DO 13 II=1,NX
      I=II
      PROD(II,J)=0.
      DO 13 K=1,NSB
   13 PROD(II,J)=PROD(II,J)+(X(K,J)/PC(J))*B(K,I)
      DO 14 I=1,NVEC
      WRITE(NSORT,11)
   14 WRITE(NSORT,9)(PROD(I,J),J=1,NA)
    9 FORMAT(3H **  10F12.8)
C*****NORMES DES FACTEURS DANS LA METRIQUE PL(I),ET CALCUL DES
C***** COSINUS AVEC LES FORMES LINEAIRES COORDONNEES
      DO 400 J=1,NFAC
      BNORM(J)=0
      DO 400 I=1,NSB
  400 BNORM(J)=BNORM(J)+PL(I)*B(I,J)*B(I,J)
      DO 401 J=1,NFAC
      DO 402 I=1,NSB
  402 B(I,J)=B(I,J)*SQRT(PL(I)/BNORM(J))
      WRITE(NSORT,11)
   11 FORMAT(1H /)
      WRITE(NSORT,403)((LX(I),B(I,J)),I=1,NSB)
  401 CONTINUE
  403 FORMAT(1H  8(A3,F8.3,4H  * ))
```

```
C***** VERIFICATIONS
      DO 501 K=1,NFAC
      TES=0
      DO 500 J=1,NA
  500 TES=TES+PC(J)*PROD(K,J)
  501 WRITE(NSORT,502)TES
  502 FORMAT(1H // 1H F15.8)
      DO 506 K=1,NFAC
      F1(K)=0
      DO 506 J=1,NA
  506 F1(K)=F1(K)+PC(J)*PROD(K,J)*PROD(K,J)
      DO 507 L=1,NSB
      F2(L)=0
      DO 507 J=1,NA
  507 F2(L)=F2(L)+(X(L,J)-PL(L)*PC(J))*(X(L,J)-PL(L)*PC(J))/PC(J)
C***** COEFFICIENTS DE CORRELATION ENTRE LES VALEURS DES FACTEURS POUR
C***** LES SOMMETS ET LES VARIABLES INITIALES
      DO 505 L=1,NSB
      DO 504 K=1,NFAC
      B(K,L)=0
      DO 503 J=1,NA
  503 B(K,L)=B(K,L)+PROD(K,J)*X(L,J)
  504 B(K,L)=B(K,L)/SQRT(F1(K)*F2(L))
  505 CONTINUE
      WRITE(NSORT,11)
      WRITE(NSORT,126)(F1(K),K=1,NFAC)
      WRITE(NSORT,11)
      WRITE(NSORT,126)(F2(L),L=1,NSB)
      WRITE(NSORT,801)
  801 FORMAT(///40H CORRELATIONS VARIABLES-INDICES                    //)
      DO 508 K=1,NFAC
      WRITE(NSORT,11)
      WRITE(NSORT,403)((LX(L),B(K,L)),L=1,NSB)
  508 CONTINUE
      NX2=4
      DO 300 J=1,NX2,2
      DO 198 I=1,NA
      F1(I)=PROD(J,I)
      F2(I)=PROD(J+1,I)
  198 CONTINUE
      NTOT=NA
C***** SORTIE GRAPHIQUE DU GRAPHE RECONSTITUE
      CALL GRAPH(F1,F2,JCAR,NTOT,60,1)
      DO 199 I=1,NSB
      F1(I)=B(J,I)
      F2(I)=B(J+1,I)
  199 CONTINUE
C*****  REPRESENTATION GRAPHIQUE DES CORRELATIONS VARIABLES-FACTEURS
      CALL GRAPH(F1,F2,LX,NSB,30,1)
  300 CONTINUE
  222 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE AFT2(X,A,B,PC,PL,NA,NSB,NO,NV,JCAR,KA,NL)
      DIMENSION B(NV,NV)
      DIMENSION PL(1),JCAR(1)
      DIMENSION X(NV,NO),A(NV,NV),PC(1)
      DIMENSION CC(100)
      DIMENSION KA(9,100)
      DIMENSION NCENT(100)
      DIMENSION VEC(50,50)
      COMMON/ENSOR/NINT,NSORT
      COMMON/NORM/DENO
C***** CALCUL DE LA MATRICE D INERTIE TOTALE
C*****    (PAR RAPPORT AU CENTRE DE GRAVITE)
C*****    B REPRESENTE LA MATRICE A DIAGONALISER SYMETRISEE
      DO 5 J=1,NSB
      DO 5 I=1,NSB
      B(I,J)=0
      RAC=SQRT(PL(I)*PL(J))
      DO 6 K=1,NA
    6 B(I,J)=B(I,J)+X(I,K)*X(J,K)/(PC(K)*RAC)
      B(I,J)=B(I,J)-RAC
      CC(I)=B(I,I)
    5 CONTINUE
C***** CALCUL DES PROFILS
      DO 111 I=1,NA
      DO 111 J=1,NSB
  111 X(J,I)=X(J,I)/PC(I)
C***** CALCUL DE LA MATRICE D INERTIE LOCALE
      DO 40 J=1,NSB
      DO 40 N=J,NSB
      AAA=0
      DO 30 I=1,NA
      PI=PC(I)
      DO 30 L=1,NL
      IF(KA(L,I))30,30,32
   32 NN=KA(L,I)
      PP=PI*PC(NN)
      AAA=AAA      +PP*(X(J,I)-X(J,NN))*(X(N,I)-X(N,NN))
   30 CONTINUE
      A(J,N)=AAA
      A(N,J)=AAA
   40 CONTINUE
      DENO=0
      DO 56 I=1,NA
      PI=PC(I)
      DO 50 L=1,NL
      NN=KA(L,I)
      IF(NN)56,56,50
   50 DENO=DENO+PI*PC(NN)
   56 CONTINUE
      DO 70 J=1,NSB
      DO 70 N=1,NSB
      A(J,N)=A(J,N)/(DENO*2.0)
   70 A(J,N)=A(J,N)/SQRT(PL(J)*PL(N))
      CALL IMPA(A,JCAR,NA,NSB,NV)
C***** CALCUL DE LA MATRICE DES CORRELATIONS LOCALES
      DO 81 I=1,NSB
      DO 81 J=1,NSB
      VEC(I,J)=A(I,J)/(SQRT(A(I,I)*A(J,J)) )
```

```
   81 CONTINUE
      CALL IMPA(VEC,JCAR,NA,NSB,NV)
      DO 71 I=1,NSB
      CC(I)=A(I,I)/CC(I)
      III=I+NA
   71 WRITE(NSORT,72)JCAR(III),CC(I)
   72 FORMAT(// 1X,A3, 5X,F8.4          )
C***** CALCUL DE LA MATRICE D INERTIE DIFFEREE,OU EXTERNE
      DO 82 I=1,NSB
      DO 82 J=1,NSB
   82 A(I,J)=(B(I,J)-A(I,J)*DENO)/(1.-DENO)
C  ATTENTION   A DESIGNE MAINTENANT  LA MATRICE EXTERNE
C***** RESTITUTION DU TABLEAU DE DONNEES X
      DO 112 I=1,NA
      DO 112 J=1,NSB
  112 X(J,I)=X(J,I)*PC(I)
      RETURN
      END
```

```
      SUBROUTINE BSPAS(B,S,VEC,NV,NSB,NFAC)
      DIMENSION B(NV,NV),S(NV,NV),VEC(NV,NV)
C*****  CALCUL DE LA CONGRUENTE DE B    =(TRANSPOSEE DE S)*B*S
      DO 121 I=1,NSB
      DO 121 J=1,NFAC
      VEC(I,J)=0
      DO 121 K=1,NSB
  121 VEC(I,J)=VEC(I,J)+B(I,K)*S(K,J)
      DO 122 I=1,NFAC
      DO 122 J=1,NFAC
      B(I,J)=0
      DO 122 K=1,NSB
  122 B(I,J)=B(I,J)+S(K,I)*VEC(K,J)
      RETURN
      END


      SUBROUTINE PROP(EIG,NSB,VEC,NV,NVEC)
      DIMENSION EIG(1),VEC(NV,NV)
      COMMON/ENSOR/NINT,NSORT
C***** EDITION DES VALEURS PROPRES,POURCENTAGES D INERTIE,VECT.PROPRES
      WRITE(NSORT,10)
   10 FORMAT(21H VALEURS PROPRES                    )
      WRITE(NSORT,9)(EIG(K),K=1,NSB)
    9 FORMAT(1H 10F12.8)
C  POURCENTAGES
      SOMV=0
      DO 700 J=1,NSB
  700 SOMV=SOMV+EIG(J)
      WRITE(NSORT,701)SOMV
  701 FORMAT(/  10H  TRACE=     F15.5)
      DO 800 J=1,NSB
      JM1=J
      POURC=(EIG(J)/SOMV)*100
  800 WRITE(NSORT,900)JM1,POURC
  900 FORMAT(/ 13H VAL.PROPRE     I3,15H  POURCENTAGE     F6.2)
      WRITE(NSORT,11)
   11 FORMAT(30H  VECTEURS PROPRES                  )
      DO 12 I=1,NVEC
   12 WRITE(NSORT,9)(VEC(J,I),J=1,NSB)
      RETURN
      END
```

II-6. Exemples d'application.

Etude de la distribution simultanée de 29 catégories d'activité dans les 80 quartiers de Paris.

II-6.1 - Nous procéderons d'abord à une analyse des correspondances du tableau de contingence croisant les deux partitions de la population parisienne en 29 catégories d'activité d'une part, et en 80 quartiers correspondant à leur résidence d'autre part.

Les données analysées sont issues du recensement de la population réalisé par l'INSEE en 1968.

Les quartiers de Paris seront repérés par trois chiffres : les deux premiers donnant le numéro de l'arrondissement, le dernier donnant le numéro du quartier dans l'arrondissement.

Les catégories d'activité seront repérées par trois lettres ou trois caractères. La correspondance avec la nomenclature complète figure ci-dessous.

Désignation de la catégorie	Symbole
Industriels	IND
Artisans	ART
Gros commerçants	GRO
Petits commerçants	PET
Professions libérales	LIB
Professeurs, prof. littéraires et scientif.	PRO
Ingénieurs	ING
Cadres administratifs supérieurs	CAS
Instituteurs, prof. intellectuelles diverses	INS
Services médicaux et sociaux	MED
Techniciens	TEC
Cadres administratifs moyens	CAM
Employés de bureau	BUR
Employés de commerce	COM
Contremaîtres	CON
Ouvriers qualifiés	QUA
Ouvriers spécialisés	SPE

Désignation de la catégorie (suite)	Symbôle
Apprentis ouvriers	APP
Manoeuvres	MAN
Gens de maison	GEN
Femmes de ménage	FEM
Autres personnels de service	AUT
Etudiants et élèves	ETU
Retirés des affaires	RAF
Retraités des services publics	RSP
Anciens salariés du secteur privé	ASR
Personnes non actives de 0 à 16 ans	GOS
Personnes non actives de 17 à 64 ans	NA2
Personnes non actives de 65 ans et plus	OLD

Les tableaux ci-après nous donnent, pour chaque catégories d'activité et pour chaque quartier de Paris : leur masse relative, leur distance (du khi-2) à l'origine (notée DISTO), leurs coordonnées sur les trois premiers axes factoriels, leurs contributions absolues et relatives à ces axes.

Les trois permiers axes représentent respectivement 64%, 13%, 7% de l'inertie totale.

La figure 1 représente les projections des deux nuages dans le plan des deux premiers facteurs.

Le premier axe, extrêmement prépondérant est un axe de statut social, où, si l'on préfère, de ségrégation sociale du lieu de résidence. Un tel axe a déjà été mis en évidence à propos de l'analyse des mêmes catégories d'activité pour toutes les communes de la région parisienne (1).

Le second axe, bien que beaucoup moins important, sépare de façon assez nette la rive droite (partie supérieure) de la rive gauche de la Seine, à quelques exceptions près, qui occupent de toutes façons des positions assez intermédiaires.

---

(1) Cf. "La morphologie sociale des communes urbaines". L. LEBART et N. TABARD. Consommation n°2 - 1971.

LES CONTR. ABS. SONT EXPRIMEES EN POURCENTAGES

NOMS	MASSES	DISTO	COORDONNEES			CONTRIBUTIONS ABSOLUES			CONTRIBUTIONS RELATIVES		
			F1	F2	F3	F1	F2	F3	F1	F2	F3
IND	.002	.3022	.3011	.2492	.0203	.25	.83	.01	.306	.205	.001
ART	.013	.1038	-.1281	.1302	-.0262	.30	1.51	.11	.158	.163	.007
GRO	.009	.2415	.2894	.2105	.0439	1.05	2.70	.21	.347	.183	.008
PET	.024	.0533	-.1420	.0347	.0012	.68	.20	.00	.378	.023	.000
LIB	.007	.5757	.6991	.1474	.0211	5.09	1.10	.04	.849	.038	.001
PRO	.014	.4073	.4625	-.3554	-.0587	4.11	11.83	.57	.525	.312	.008
ING	.011	.2658	.4263	-.1693	.1174	2.73	2.09	1.78	.684	.108	.052
CAS	.030	.2451	.4638	.0496	.1219	9.22	.51	5.48	.878	.010	.061
INS	.014	.0915	.2327	-.1059	.0682	1.11	1.12	.82	.592	.123	.051
MED	.007	.0823	.0849	-.1719	.1036	.07	1.41	.91	.088	.359	.130
TEC	.019	.0580	-.1363	-.1345	.1033	.50	2.38	2.49	.320	.312	.184
CAM	.037	.0336	.0248	-.0734	.1406	.03	1.36	8.87	.018	.160	.588
BUR	.103	.0342	-.1578	-.0378	.0508	3.64	1.01	3.25	.729	.042	.076
COM	.022	.0491	-.1464	.0261	.0305	.67	.10	.25	.437	.014	.019
CON	.007	.1435	-.3112	-.0674	.0502	.99	.23	.22	.675	.032	.018
GUA	.063	.1271	-.3393	.0394	-.0707	10.29	.67	3.85	.906	.012	.039
SPE	.043	.1578	-.3642	.0758	-.1169	8.16	1.71	7.24	.841	.036	.087
APP	.002	.2459	-.3527	.0039	-.0653	.29	.00	.09	.506	.000	.017
MAN	.025	.2627	-.4242	.1193	-.2035	6.41	2.46	12.71	.685	.054	.158
GEN	.015	1.3433	.9875	.5680	-.0424	21.38	34.32	.34	.726	.240	.001
FEM	.011	.0805	.0613	.1913	-.0692	.06	2.71	.63	.047	.455	.059
AUT	.043	.0458	-.0194	.0866	.0112	.02	2.23	.07	.008	.164	.003
ETU	.059	.3055	.4466	-.2022	-.2423	16.61	16.51	42.09	.653	.134	.192
RAF	.008	.0610	.0648	.0414	.0760	.05	.10	.57	.069	.028	.095
RSP	.023	.1029	.0189	-.2465	.0975	.01	9.47	2.63	.003	.590	.092
ASP	.079	.0284	-.1290	-.0191	.0526	1.86	.20	2.66	.586	.013	.097
GOS	.160	.0109	-.0466	.0071	-.0002	.49	.06	.00	.199	.005	.000
NA2	.114	.0152	.0480	.0331	.0080	.37	.86	.09	.151	.072	.004
CLD	.038	.0789	.2547	.0345	.0656	3.53	.31	2.01	.822	.015	.054

NOMS	MASSES	DISTO	COORDONNEES			CONTRIBUTIONS ABSOLUES			CONTRIBUTIONS RELATIVES		
			F1	F2	F3	F1	F2	F3	F1	F2	F3
011	.001	.0930	.0802	.0790	-.0699	.01	.05	.07	.069	.067	.053
012	.007	.1280	-.1546	.1112	-.0551	.23	.58	.25	.187	.097	.024
013	.002	.1054	.1310	.0700	.0796	.05	.07	.17	.163	.047	.060
014	.002	.1441	.1332	.1189	.0115	.06	.22	.00	.123	.098	.001
021	.001	.1544	.0853	.0475	.0083	.01	.02	.00	.047	.015	.000
022	.002	.0830	-.1092	.0655	.0461	.03	.06	.05	.144	.052	.026
023	.004	.1082	-.2084	.1280	-.0260	.25	.45	.03	.401	.151	.006
024	.007	.1540	-.2758	.1738	-.1532	.76	1.47	2.03	.494	.196	.152
031	.006	.1981	-.2925	.1836	-.1713	.76	1.46	2.25	.432	.170	.148
032	.006	.0823	-.1866	.1061	-.0639	.28	.43	.28	.423	.137	.050
033	.005	.0548	-.1577	.0757	-.0453	.18	.20	.13	.453	.104	.037
034	.005	.1215	-.2668	.1077	-.1286	.46	.37	.93	.586	.096	.136
041	.005	.0729	-.1681	.1149	-.0816	.19	.43	.39	.387	.181	.091
042	.008	.0479	-.1552	.0570	-.0950	.26	.17	.84	.503	.068	.188
043	.006	.0272	.0491	-.0554	-.0592	.02	.12	.25	.089	.113	.129
044	.003	.1374	.2672	-.0707	.0268	.27	.09	.02	.520	.036	.005
051	.007	.0650	.0794	-.1773	-.0291	.06	1.51	.07	.097	.484	.013
052	.009	.0538	.0124	-.1677	.0848	.00	1.80	.82	.003	.522	.133
053	.010	.1608	.2399	-.2456	-.1027	.78	3.97	1.23	.358	.375	.066
054	.006	.0956	.1726	-.1481	-.1461	.24	.86	1.49	.312	.229	.223
061	.004	.0897	.1718	-.1005	-.0559	.18	.29	.16	.329	.113	.035
062	.006	.2268	.4000	-.1738	-.1405	1.26	1.15	1.34	.706	.133	.087
063	.014	.2016	.3903	-.1748	-.1104	2.96	2.88	2.04	.756	.152	.060
064	.004	.1316	.3038	-.0982	-.0497	.47	.24	.11	.702	.073	.019
071	.008	.3106	.5301	.0243	-.0972	3.16	.03	.92	.905	.002	.030
072	.004	.2699	.4911	.0865	-.0238	1.32	.20	.03	.894	.028	.002
073	.007	.1903	.4153	-.0592	-.0097	1.69	.17	.01	.906	.018	.000
074	.015	.1081	.3136	.0402	.0144	2.12	.17	.04	.909	.015	.002
151	.034	.0460	-.0022	-.1374	.1486	.00	4.47	9.30	.000	.410	.480
152	.020	.0237	.0605	-.1052	.0436	.10	1.54	.47	.154	.467	.080
153	.019	.0221	.0683	-.0706	.0259	.13	.67	.16	.211	.226	.030
154	.016	.0427	.0483	-.1017	.0545	.05	1.15	.59	.055	.242	.070
161	.033	.1975	.4141	-.0370	.0667	8.08	.31	1.81	.868	.007	.023
162	.023	.4178	.6024	.1996	-.0211	11.81	6.29	.12	.868	.095	.001
163	.015	.5118	.6235	.3162	-.0708	8.10	10.11	.90	.760	.195	.010
164	.012	.3198	.4850	.2728	-.0061	4.10	6.29	.01	.735	.233	.000
081	.004	.3824	.4959	.3199	-.0145	1.36	2.75	.01	.643	.268	.001
082	.006	.2630	.4414	.2344	-.0091	1.78	2.43	.01	.741	.209	.000
083	.005	.1690	.2871	.2123	.0155	.53	1.42	.01	.458	.267	.001
084	.011	.1740	.3798	.0895	-.0213	2.27	.61	.06	.829	.046	.003
091	.012	.0470	.1326	.0122	.0591	.30	.01	.52	.374	.003	.074
092	.003	.1187	.0431	.0988	.0669	.01	.20	.16	.016	.082	.038
093	.006	.0880	-.0261	.1147	.0082	.01	.52	.03	.008	.150	.001
094	.012	.0289	.0359	-.0038	.0765	.02	.00	.87	.045	.001	.202
171	.019	.0987	.2772	.0531	.1010	2.12	.38	2.43	.779	.029	.103
172	.019	.1898	.4167	.1011	.0228	4.76	1.36	.12	.915	.054	.003
173	.019	.0244	-.0184	-.0052	.1048	.01	.00	2.56	.014	.001	.451
174	.023	.0465	-.1924	-.0163	.0383	1.23	.04	.42	.795	.006	.032
101	.012	.0192	-.0193	-.0376	.0551	.01	.11	.43	.019	.073	.158
102	.008	.0605	-.1136	.0901	-.0085	.15	.46	.01	.213	.134	.001
103	.012	.0423	-.1403	.0690	-.0139	.34	.39	.03	.465	.113	.005
104	.013	.0872	-.2753	.0567	-.0496	1.36	.28	.38	.869	.037	.028
181	.034	.0363	-.1394	-.0081	.0934	.93	.02	3.61	.536	.002	.240
182	.036	.0467	-.1834	.0196	.0578	1.71	.09	1.46	.720	.008	.072
183	.014	.1336	-.3295	.0893	-.0826	2.15	.77	1.16	.812	.060	.051
184	.009	.1707	-.3618	.0806	-.1077	1.61	.39	1.23	.767	.038	.068
191	.017	.1531	-.3237	.0849	-.1309	2.60	.87	3.67	.684	.047	.112
192	.006	.1692	-.3715	.0934	-.0790	1.09	.33	.42	.816	.052	.037
193	.019	.0373	-.1660	-.0052	.0189	.74	.00	.08	.738	.001	.010
194	.016	.0505	-.2086	.0143	-.0001	.99	.02	.00	.861	.004	.000
111	.017	.1043	-.2716	.1174	-.0867	1.75	1.59	1.54	.707	.132	.072
112	.014	.0509	-.1918	.0486	-.0113	.73	.23	.02	.723	.046	.003
113	.023	.0765	-.2599	.0592	-.0252	2.19	.55	.18	.852	.046	.008
114	.017	.0586	-.2177	.0325	-.0301	1.16	.13	.19	.809	.018	.015
121	.015	.0442	.0373	-.1371	.1232	.03	1.90	2.72	.031	.425	.343
122	.027	.0238	-.0858	-.0762	.0638	.29	1.10	1.37	.309	.243	.171
123	.004	.0676	-.1673	-.0574	.0434	.14	.08	.08	.414	.349	.028
124	.016	.0314	-.1392	.0004	-.0718	.43	.00	.98	.617	.000	.164
201	.016	.1493	-.3285	.1034	-.1209	2.83	1.22	2.95	.723	.072	.098
202	.014	.0368	-.1223	-.0479	.0284	.31	.23	.14	.406	.062	.022
203	.018	.0853	-.2780	.0452	-.0042	2.03	.26	.00	.905	.024	.000
204	.024	.0734	-.2462	.0210	-.0311	2.07	.07	.28	.826	.006	.013
131	.008	.0572	-.0913	-.1680	-.0020	.09	1.55	.00	.146	.493	.000
132	.020	.0879	-.2350	-.0109	-.0441	1.53	.02	.47	.628	.001	.022
133	.025	.0320	-.1213	-.0777	.0069	.53	1.06	.01	.460	.189	.001
134	.008	.0417	.0362	-.1503	.0506	.02	1.31	.26	.031	.541	.061
141	.010	.0983	.1650	-.1211	.0214	.40	1.04	.06	.277	.149	.005
142	.008	1.1192	.5469	-.5921	-.6341	3.34	18.97	36.62	.267	.313	.359
143	.019	.0605	.1215	-.1856	.0687	.39	4.40	1.07	.244	.569	.078
144	.028	.0309	-.1953	-.0526	.0158	1.52	.54	.09	.750	.054	.005

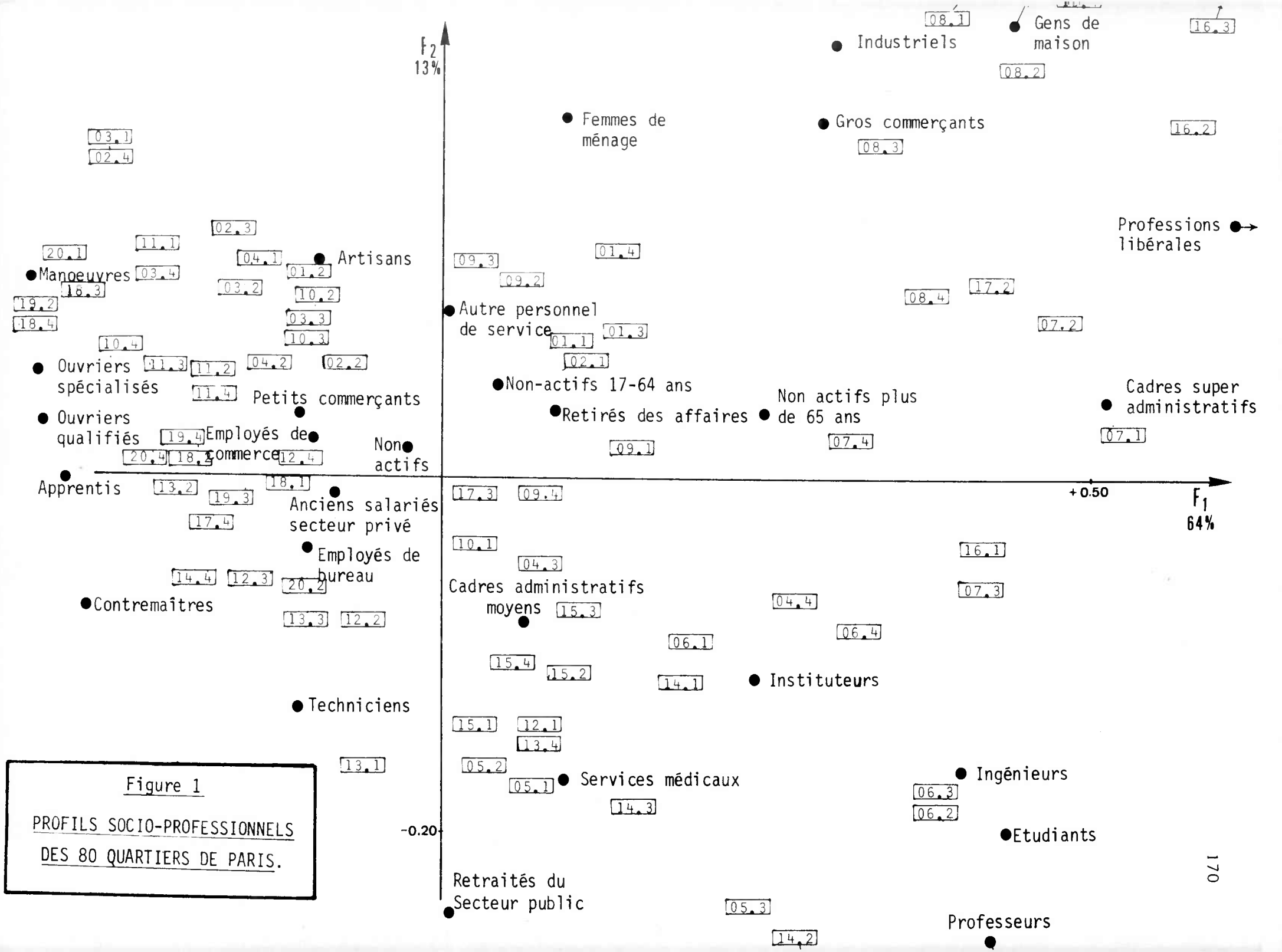


Figure 1

PROFILS SOCIO-PROFESSIONNELS DES 80 QUARTIERS DE PARIS.

II-6.2 - Nous avons ensuite substitué, dans cette analyse, la matrice d'inertie locale à la matrice d'inertie totale en considérant ces tableaux comme des matrices de similarités différentes. La nouvelle typologie des catégories d'activité est représentée sur la figure 2.

On peut faire les remarques suivantes :

a) Dans l'ensemble, l'essentiel de la typologie est conservé, l'axe dominant est toujours un axe de statut social. Ce résultat n'était pas évident "a priori", car s'il est clair qu'il existe d'importantes zones où la population est aisée et d'autres où elle occupe un statut social modeste, phénomène qui se traduit pas une cohabitation préférentielle de certains types de professions, il est assez surprenant que ce phénomène soit très sensible, disons à l'intérieur des arrondissements. Bien entendu, le phénomène est moins typé (les deux premiers facteurs représentent respectivement 47% et 16% de l'inertie locale totale) ; l'essentiel de la structure subsiste néanmoins.

b) Cependant, certaines modifications peuvent être interprétées : le point "Artisans", dans le quadrant supérieur gauche de la figure 1 rejoint l'origine dans la figure 2. Autrement dit, les artisans habitent électivement une certaine partie de Paris, ce qui les associe à certaines catégories de population. Cependant, localement, il ne semblent pas avoir d'atavisme particulier.

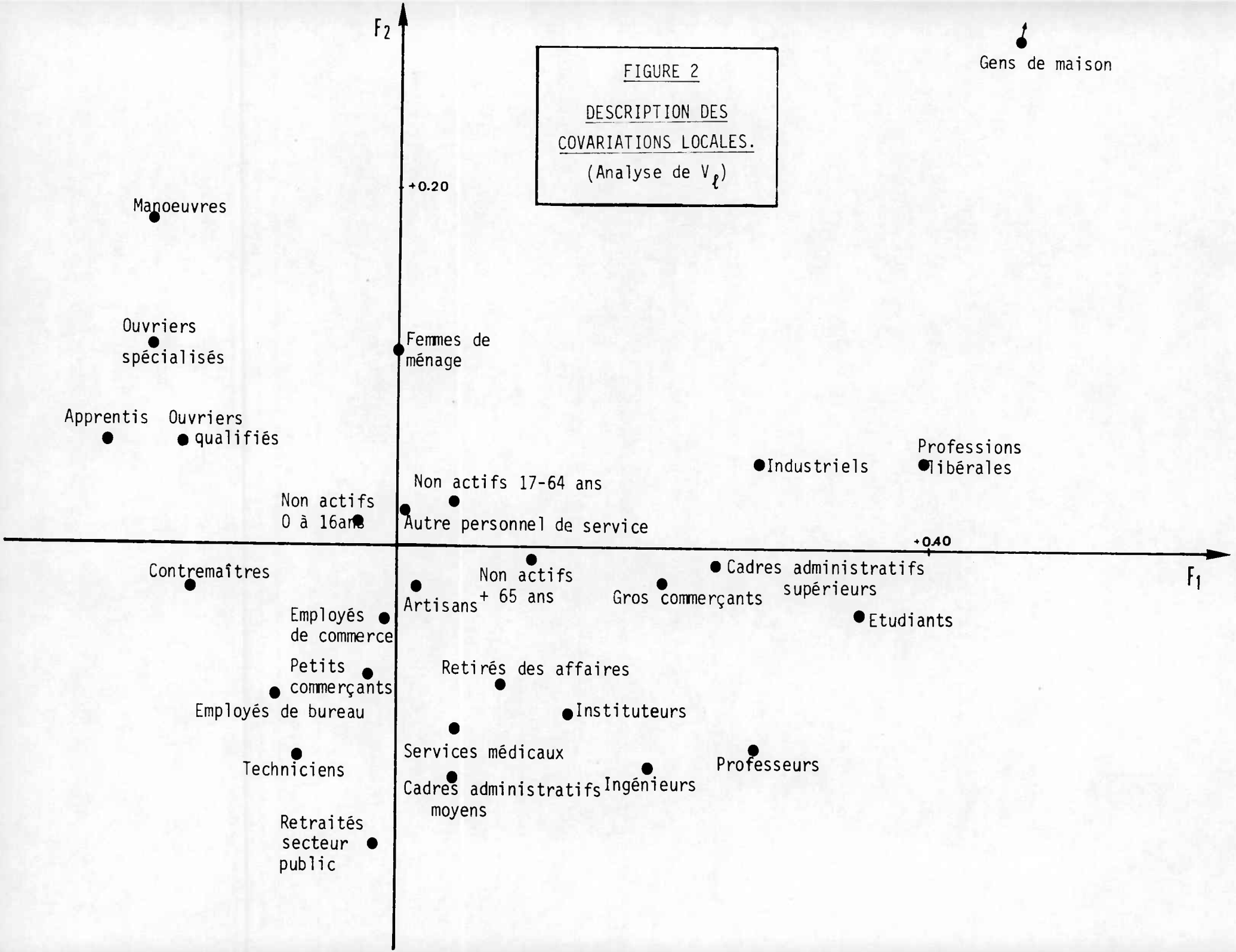
De façon analogue, les points "Industriels" et "Gros commerçants" du quadrant supérieur droit de la figure 1 descendent près de l'axe des abscisses de la figure 2, où ils sont presque rejoints par le point "Etudiants", en provenance du quadrant inférieur gauche de la figure 1.

Ainsi, ces catégories, séparées dans le bilan global que constitue l'analyse des correspondances à cause de la vocation particulière de certaines grandes zones, sont en réalité assez étroitement associées.

FIGURE 2

DESCRIPTION DES COVARIATIONS LOCALES.

(Analyse de V_ℓ)



II-6.3 - Recherche des indices descriptifs de contiguïté minimale.

Ce calcul, qui revient à décomposer simultanément les formes d'inertie locale et totale, permet de donner une vue synthétique de la compatibilité existant entre la structure de graphe et la description statistique des sommets du graphe.

La figure 3 nous montre quelle carte on peut espérer reconstituer avec des coordonnées socio-professionnelles. Les zones encerclées représentent des arrondissements.

Notons que les coefficients de contiguïté des deux premiers indices (ou facteurs) valent respectivement 0.071 et 0.109. La même analyse faite avec les mêmes données statistiques et un graphe dont les arêtes sont prises au hasard nous fournit des facteurs dont les coefficients de contiguïté valent respectivement 0.63 et 0.69.

COEFFICIENTS DE CORRELATION ENTRE LES VALEURS DU PREMIER INDICE ET LES VARIABLES

IND	.192	ART	-.659	GRO	.200	PET	-.702
INS	.633	MED	.461	TEC	-.158	CAM	.351
SPE	-.748	APP	-.525	MAN	-.704	GEN	.594
RSP	.328	ASR	-.524	GOS	-.298	NA2	.399
LIB	.609	PRO	.604	ING	.850	CAS	.790
BUR	-.580	COM	-.659	CON	-.508	QUA	-.793
FEM	.099	AUT	-.212	ETU	.547	RAF	.044
OLD	.755						

COEFFICIENTS DE CORRELATION ENTRE LES VALEURS DU SECOND INDICE ET LES VARIABLES

IND	-.428	ART	-.431	GRO	-.593	PET	-.180
INS	.149	MED	.447	TEC	.472	CAM	-.014
SPE	.101	APP	.162	MAN	.103	GEN	-.408
RSP	.704	ASR	-.013	GOS	.260	NA2	-.168
LIB	-.297	PRO	.468	ING	.088	CAS	-.255
BUR	.165	COM	-.209	CON	.323	QUA	.147
FEM	-.624	AUT	-.490	ETU	.146	RAF	-.421
OLD	-.316						

Les deux tableaux ci-dessus nous donnent des quantités s'interprétant de façon analogue aux contributions relatives de l'analyse des correspondances. On peut noter la profonde ressemblance entre le premier indice et le premier facteur de l'analyse des correspondances précédente ; au signe près, le second indice ressemble également au second facteur issu de cette analyse.

Sur la figure 3, le trait gras est censé représenter la Seine : le seul quartier ne figurant pas sur la bonne rive est le quartier encadré : 04-4 (Ile Saint-Louis, officiellement classée rive droite, ce qui est démenti ici pas son profil socio-professionnel).

Ce que nous appelons indice descriptif est une forme linéaire pouvant être représentée par un vecteur à p composantes, si p désigne le nombre de catégories d'activité.

Les valeurs de ces composantes sont délicates d'interprétation, d'autant plus qu'elles sont plus instables que les coefficients de corrélation variable s-facteurs, par suite des colinéarités existant entre les variables. Cependant, leur signe présente parfois une certaine stabilité (notamment lorsqu'il résiste à de sévères troncatures du support du nuage initial, i.e. si l'on travaille seulement dans l'espace des premiers axes principaux d'inertie du nuage).

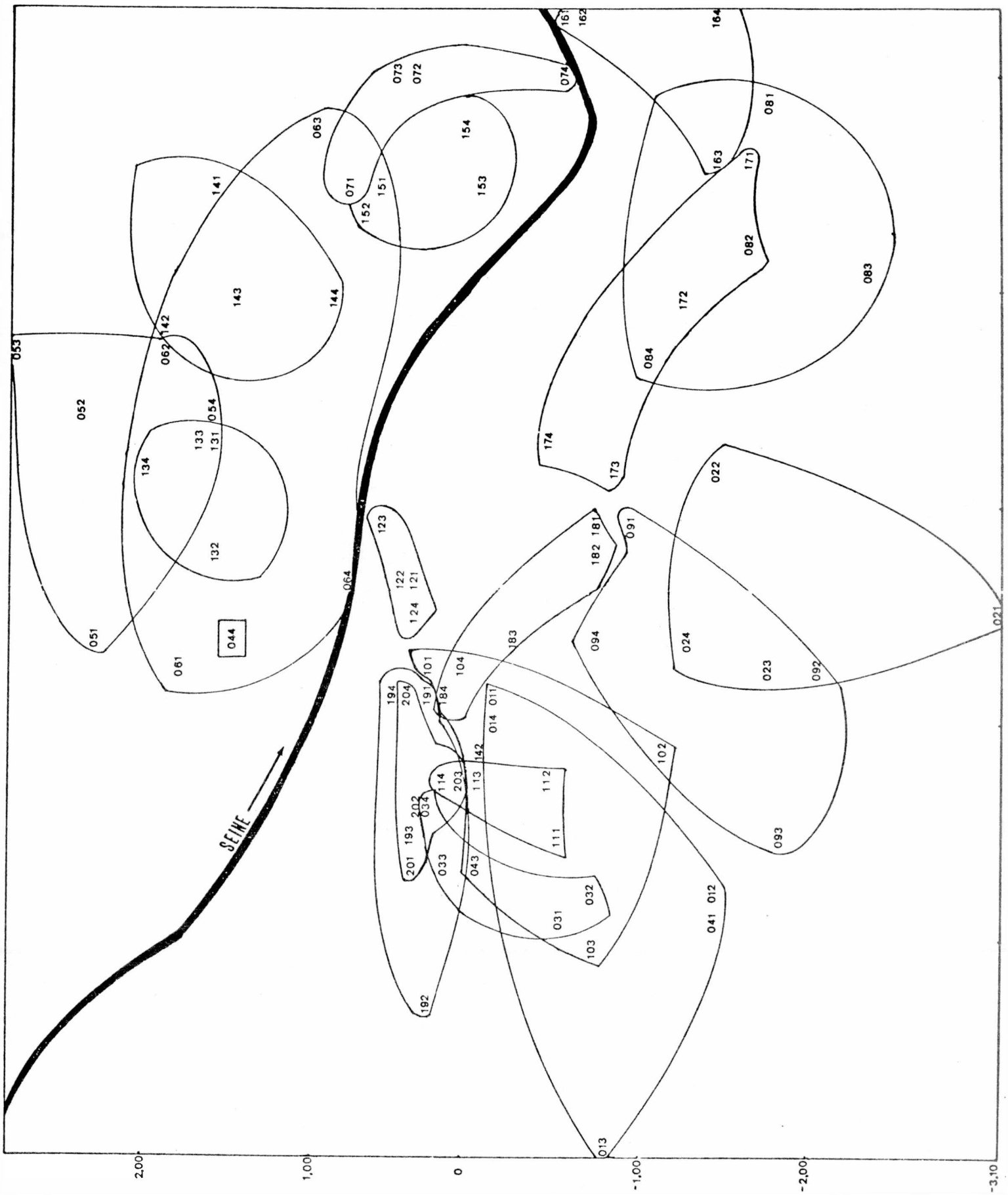
Donnons les composantes des deux premiers indices descriptifs extraits : (au lieu des composantes brutes, nous donnons en fait les cosinus de l'angle entre la forme linéaire-indice et les formes linéaires-coordonnées représentant les variables, calculés ici dans la métrique usuelle des facteurs en analyse des correspondances).

PREMIER INDICE - COEFFICIENTS

IND	-.235	ART	-.149	GRO	-.227	PET	.086
INS	-.078	MED	-.153	TEC	.485	CAM	.085
SPE	-.061	APP	.089	MAN	.068	GEN	.060
RSP	.167	ASR	-.123	GOS	-.088	NA2	.017
LIB	-.136	PRO	-.164	INC	.385	CAS	-.031
BUR	-.238	COM	-.111	CON	-.234	QUA	-.008
FEM	.196	AUT	.253	ETU	-.002	RAF	-.139
OLD	.283						

FIGURE 3

RECONSTITUTION DE LA CARTE DES 80 QUARTIERS DE PARIS
A PARTIR DE LEURS PROFILS SOCIO-PROFESSIONNELS



DEUXIEME INDICE - COEFFICIENTS

IND	.104	ART	.021	GRO	-.043	PET	.010
INS	.165	MED	.201	TEC	.363	CAM	-.393
SPE	.029	APP	.070	MAN	.015	GEN	.066
RSP	.197	ASR	-.195	GOS	.139	NA2	.031
LIB	.138	PRO	.381	ING	.124	CAS	-.319
BUR	.047	COM	-.179	CON	.048	QUA	.097
FEM	-.407	AUT	-.132	ETU	-.061	RAF	-.036
OLD	-.099						

On constate que le premier indice, par exemple, donne des poids négatifs aux catégories : Industriels, Gros commerçants, Instituteurs, Services médicaux, Professions libérales, Professeurs, alors qu'il est corrélé positivement, et même parfois fortement à toutes ces variables.

Ces signes négatifs sont conservés lorsque l'on passe des 28 dimensions initiales à 14 dimensions, en abandonnant la moitié des axes d'inertie du nuage initial.

Ce premier indice doit s'efforcer d'avoir une variance locale minimale, alors que sa variance totale est maintenue constante.

Pour ce faire, il a intérêt à opposer entre elles des variables liées localement et peu liées globalement, car cela ne peut que diminuer ses variations locales, sans porter atteinte à ces variations globales.

Les variables qui, tout en étant corrélées de la même façon à l'indice correspondent à des coefficients de signes contraires, pourraient être dans cette situation, puisque nous avons vu précédemment que les industriels, les gros commerçants, les professions libérales s'associaient aux autres variables de façon différente selon l'échelle géographique.

En fait, une de nos prochaines étapes de travail consistéra à mettre au point des critères de validation de ces coefficients.

REFERENCES BIBLIOGRAPHIQUES

DU CHAPITRE III

1) BENZECRI J.P. (1967) - Sur l'analyse de la correspondance définie par un graphe. (in "L'Analyse des Données" Tome 2 - DUNOD 1973)

2) BENZECRI J.P. (1972) - Analyse factorielle et analyse discriminante (in "L'Analyse des données" Tome 2 - DUNOD 1973)

3) BENZECRI J.P. (1969) - Sur l'analyse d'une correspondance symétrique (in "L'Analyse des Données" Tome 2 - DUNOD 1973)

4) BERGE Claude (1970) - Graphes et Hypergraphes - DUNOD 1970

5) GEARY R.C. (1954) - The contiguïty ratio and statistical mapping "The incorporated statistician" Vol. 5 - n°3 - pp 115 - 145

6) LEBART L. (1966) - "Analyse statistique de la Contiguïté" Thèse 3ème Cycle (Extraits in : Publication de l'I.S.U.P. - 1969)

CHAPITRE IV

TROIS EXEMPLES D'APPLICATION

EXEMPLE I

I - ATTITUDES PAR RAPPORT A LA POLITIQUE DES PRESTATIONS FAMILIALES

Cet exemple est tiré des résultats de l'enquête CNAF 1971. C'est un exemple d'analyse de questionnaires par la méthode des correspondances multiples décrite au chapitre II . Le tableau que l'on cherche à décrire est l'ensemble des réponses de n individus à q questions mises sous forme disjonctive .

Particularités du tableau : le tableau se présente comme une succession de q sous-tableaux disjoints (autant que de questions, il y en a 29), chacun comportant 1383 lignes (le nombre d'individus interrogés). Les variables sont les modalités des questions, leur nombre varie entre 2 et 5 selon les questions ; au total il y en a 88. Le tableau initial a donc 1383 lignes et 88 colonnes.(1).

Le tableau n'est composé que de 0 ou de 1, le 1 correspondant à la modalité de réponse choisie par l'individu considéré. Il doit y avoir une réponse et une seule par question et par individu. Les individus ont donc le même poids. Il n'en est pas de même pour les variables qui ont des poids différents.

On n'a tenu compte que des réponses exprimées. Comme l'échantillon doit figurer exhaustivement pour chaque question, les non-réponses ont été réparties aléatoirement entre les modalités prévues.

Particularité de l'information disponible : il s'agit de données d'enquêtes. L'analyse porte sur un seul thème, tandis que l'information disponible sur chaque individu est beaucoup plus vaste. L'interprétation des proximités obtenue sera enrichie par la projection, en variables illustratives, de catégories d'appartenances, de comportement... ou d'attitudes par rapport à d'autres thèmes.

---

(1) Nous avons vu au chapitre II que la description d'un tel tableau conduisait à diagonaliser une matrice d'ordre (59 x 59), puisque 59 = 88 - 29.

I-1. Thèmes contenus dans les questions.

Les questions posées couvrent plusieurs thèmes différents, relevant tous de la politique des prestations :

- la sélectivité des prestations, c'est-à-dire l'institution d'un critère de ressources pour l'obtention ou le calcul d'allocations particulières (ce qui existe pour l'allocation logement et depuis peu pour l'allocation de salaire unique ou celle de frais de garde).

- l'arbitrage entre prestations en nature (réduction de tarifs d'utilisation de services ou équipements divers, construction d'équipement) et prestations en espèces affectées (exemple allocation logement) ou non affectées (type allocations familiales proprement dites).

- l'importance relative des fonctions à assumer : logement, équipement ménager, garde, vacances, santé, services d'information ou services culturels...

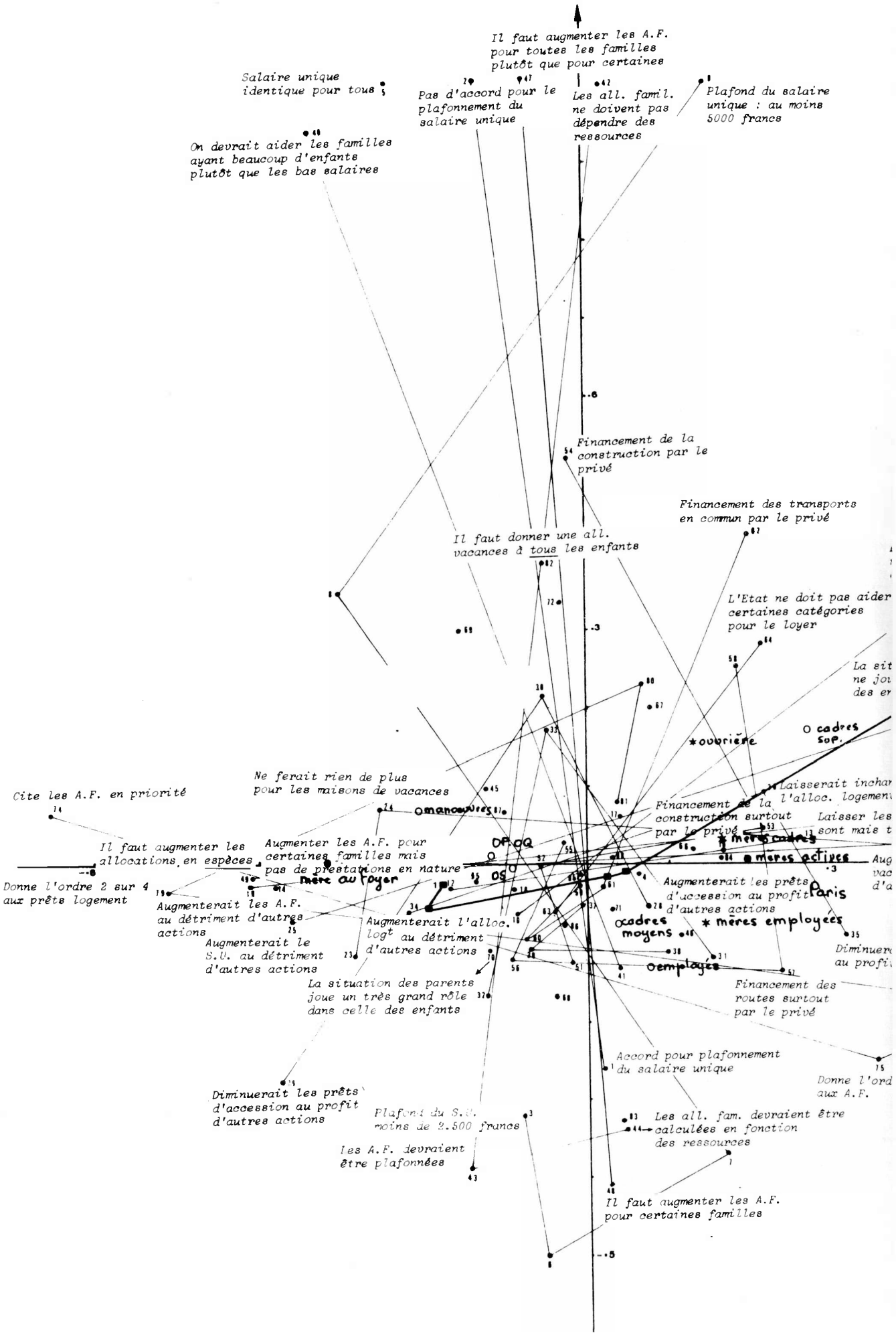
- l'arbitrage entre financement public et financement privé pour certaines actions.

La liste des questions retenues pour l'analyse est donnée à l'annexe 1 . On pourra s'y reporter en particulier pour les variables dont, faute de place, on n'a indiqué que le numéro sur les graphiques.

I-2. Signification des réponses d'après leur proximité

Les résultats figurent sur le graphique 1 (réponse des femmes) et 2 (réponse des hommes). La position relative des variables est voisine dans les deux cas.

Le premier axe oppose les attitudes en faveur des allocations en espèces (49), à gauche, aux attitudes en faveur des prestations en nature (50) - question posée en général -. Plus précisément on trouve groupées les réponses en faveur d'une augmentation des allocations familiales (19), de l'allocation de salaire unique (25), c'est-à-dire traduisant le plus grand prix attaché aux prestations en espèces non affectées, à proximité des réponses dévalorisant les prêts pour le logement (79), les prêts d'accession à la propriété (29), la construction de maisons de vacances (23). Toutes ces variables occupent la partie extrême gauche.


Il faut augmenter les A.F. pour toutes les familles plutôt que pour certaines
Salaire unique identique pour tous
Pas d'accord pour le plafonnement du salaire unique
Les all. famil. ne doivent pas dépendre des ressources
Plafond du salaire unique : au moins 5000 francs
On devrait aider les familles ayant beaucoup d'enfants plutôt que les bas salaires
Financement de la construction par le privé
Financement des transports en commun par le privé
Il faut donner une all. vacances à tous les enfants
L'Etat ne doit pas aider certaines catégories pour le loyer
ouvrière
cadres sup.
Ne ferait rien de plus pour les maisons de vacances
manoeuvres
Cite les A.F. en priorité
Financement de la construction surtout par le privé
mères cadres
mères actives
Il faut augmenter les allocations en espèces
Augmenter les A.F. pour certaines familles mais pas de prestations en nature
mère au foyer
Donne l'ordre 2 sur 4 aux prêts logement
Augmenterait les A.F. au détriment d'autres actions
Augmenterait les prêts d'accession au profit d'autres actions
Paris
cadres moyens
mères employées
Augmenterait le S.U. au détriment d'autres actions
Augmenterait l'alloc. logt au détriment d'autres actions
employées
La situation des parents joue un très grand rôle dans celle des enfants
Financement des routes surtout par le privé
Accord pour plafonnement du salaire unique
Diminuerait les prêts d'accession au profit d'autres actions
Plafond du S.U. moins de 2.500 francs
Les all. fam. devraient être calculées en fonction des ressources
les A.F. devraient être plafonnées
Il faut augmenter les A.F. pour certaines familles

Graphique 1

ATTITUDES DES FEMMES PAR RAPPORT
AUX PRESTATIONS FAMILIALES

1383 personnes - 88 variables - 29 questions

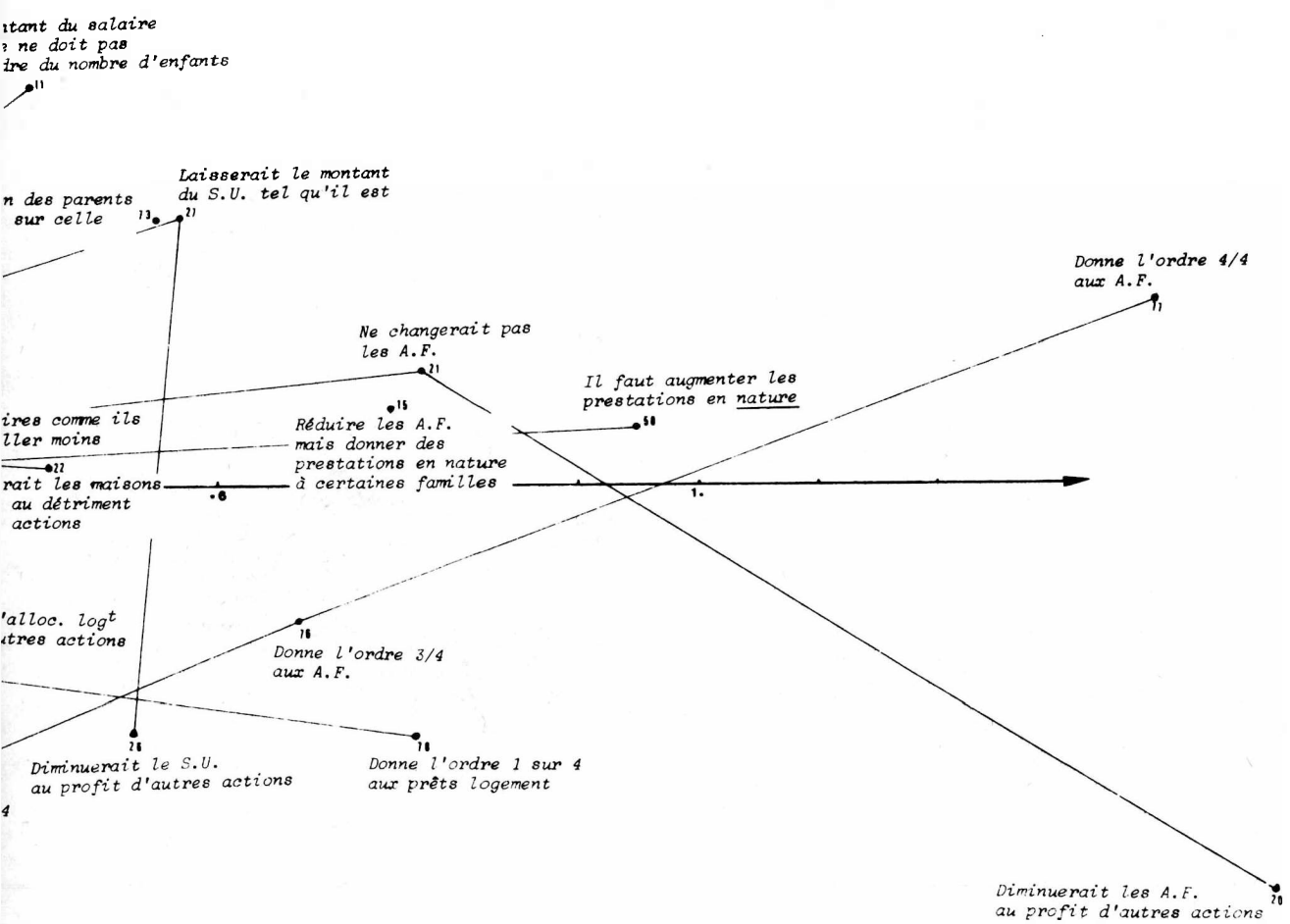
en italique : les 88 variables analysées

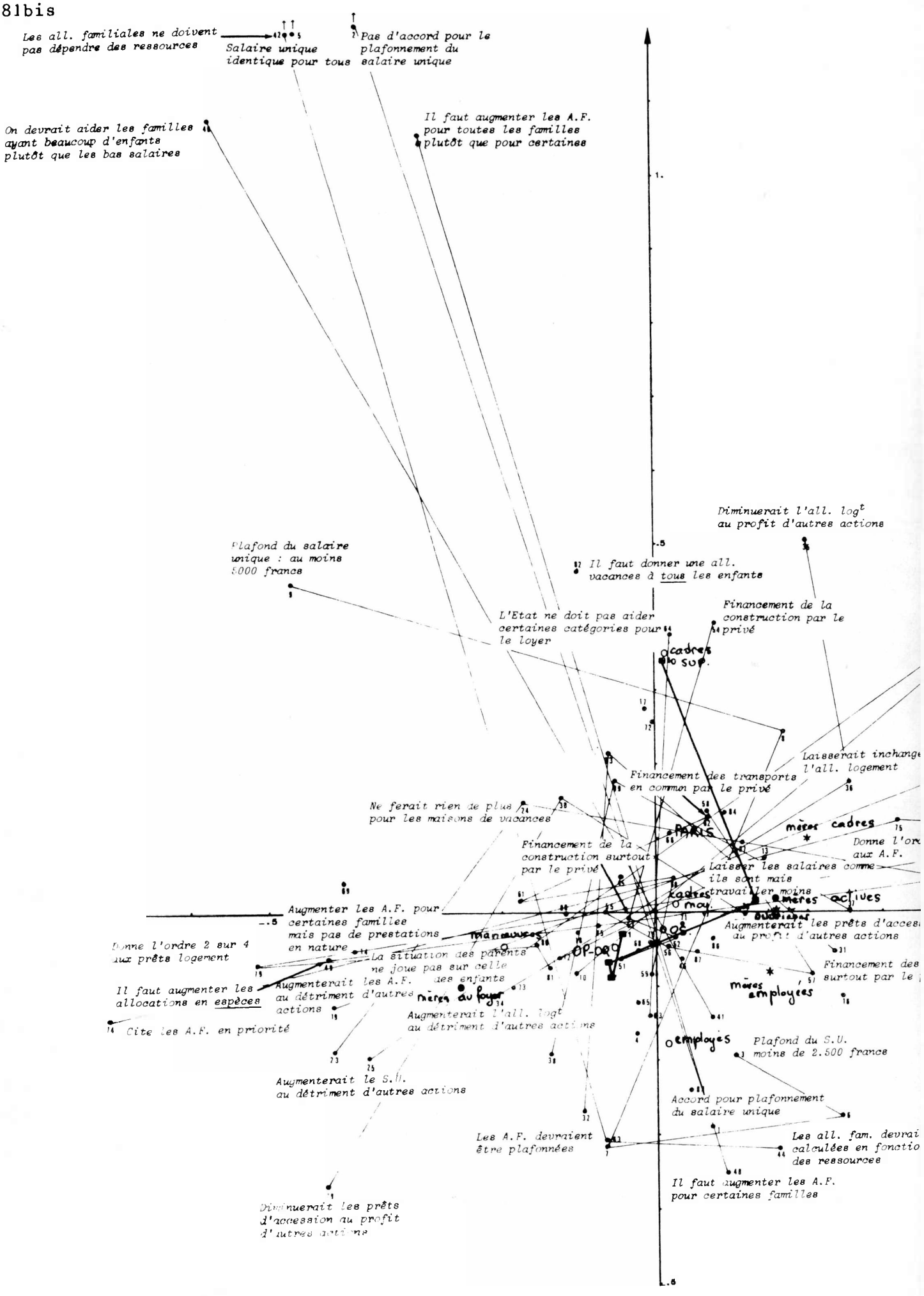
en caractère droit : quelques variables illustratives

en trait fort : le revenu, croissant de 1 à 10

o Profession du père

* Profession de la mère (exercée actuellement)



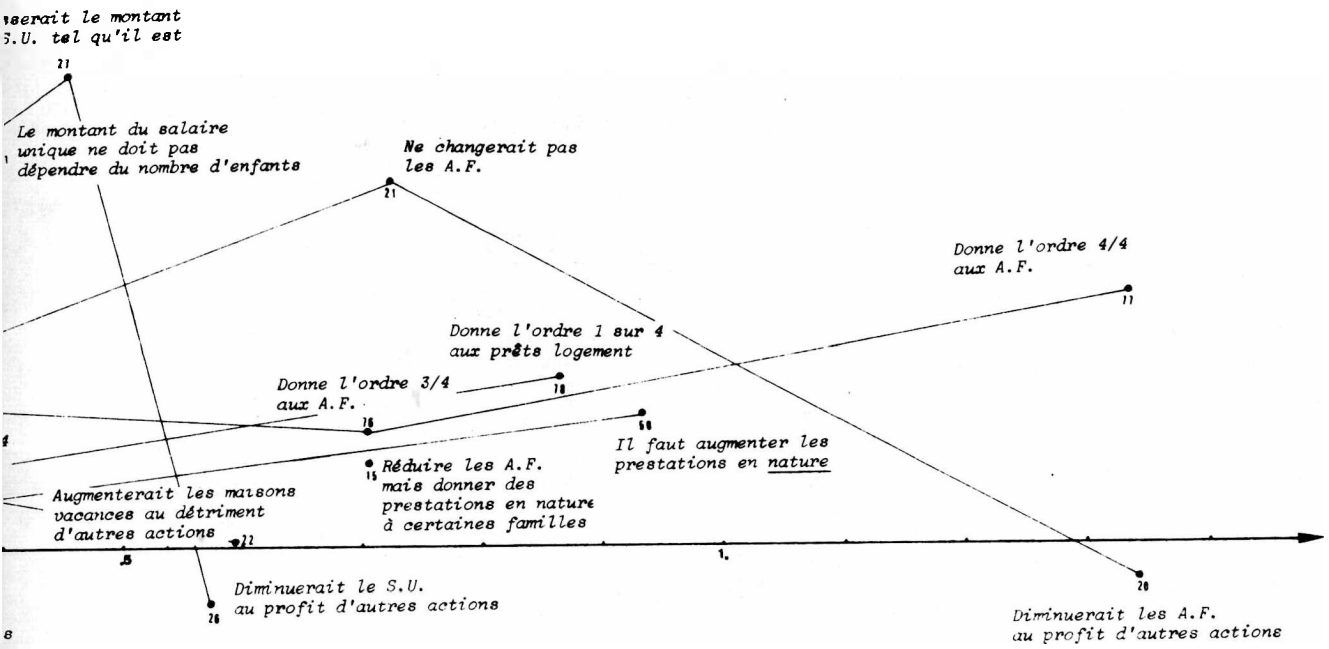

Les all. familiales ne doivent pas dépendre des ressources
Salaire unique identique pour tous
Pas d'accord pour le plafonnement du salaire unique
On devrait aider les familles ayant beaucoup d'enfants plutôt que les bas salaires
Il faut augmenter les A.F. pour toutes les familles plutôt que pour certaines
Diminuerait l'all. logt au profit d'autres actions
Plafond du salaire unique : au moins 5000 francs
Il faut donner une all. vacances à tous les enfants
L'Etat ne doit pas aider certaines catégories pour le loyer
Financement de la construction par le privé
Laisserait inchangé l'all. logement
Financement des transports en commun par le privé
Ne ferait rien de plus pour les maisons de vacances
Financement de la construction surtout par le privé
Donne l'ordre aux A.F.
Laisser les salaires comme ils sont mais travailler moins
Augmenter les A.F. pour certaines familles mais pas de prestations en nature
Augmenterait les prêts d'acces au profit d'autres actions
Donne l'ordre 2 sur 4 aux prêts logement
La situation des parents ne joue pas sur celle des enfants
Financement des surtout par le
Il faut augmenter les allocations en espèces
Augmenterait les A.F. au détriment d'autres actions
Augmenterait l'all. logt au détriment d'autres actions
Cite les A.F. en priorité
Plafond du S.U. moins de 2.500 francs
Augmenterait le S.U. au détriment d'autres actions
Accord pour plafonnement du salaire unique
Les A.F. devraient être plafonnées
Les all. fam. devrai calculées en fonctio des ressources
Il faut augmenter les A.F. pour certaines familles
Diminuerait les prêts d'accession au profit d'autres actions
cadres su.
mères cadres
PARIS
cadres moy.
mères actives
mères employées
employés
mères au foyer
manœuvres
OP-pro
1.
.5
-.5
-1.

Graphique 2

ATTITUDES DES HOMMES PAR RAPPORT
AUX PRESTATIONS FAMILIALES

1383 personnes - 88 variables - 29 questions
en italique : les 88 variables analysées
en caractère droit : quelques variables illustratives
en trait fort : le revenu, croissant de 1 à 10

o Profession du père
* Profession de la mère (enquête actuellement)



A l'opposé (partie droite) se situent les réponses en faveur des prestations en nature (50) dont la signification se précise par leur proximité avec les réponses donnant la priorité aux prêts de logement (78), la dernière place aux allocations familiales (77), les réponses diminuant les allocations familiales (20) ou le salaire unique (26) au profit d'autres formes d'actions et,à un moindre titre,celles diminuant l'allocation de logement (35). L'absence d'intérêt pour les prêts d'équipement ménager (38) se situe également du côté droit.

Le contenu de ces deux attitudes contraires se précise encore par leur proximité avec des variables ne touchant pas directement à la politique familiale : du côté gauche le choix de l'augmentation des salaires sans réduction du temps de travail (12) (pour une répartition des gains de productivité), du côté droit le choix d'une augmentation des temps de loisirs (13).

Ces résultats sont valables aussi bien en ce qui concerne les réponses des hommes que celles des femmes. L'ordre des variables le long du premier axe est approximativement le même dans les deux cas. Il oppose les actions visant à un accroissement direct des ressources des familles aux actions concernant les prêts logement (prestations affectées) ou l'accès aux services collectifs.

Pour rendre plus concrète la proximité entre certaines variables des graphiques, le tableau ci-dessous présente les fréquences relatives observées pour certains croisements entre les variables 74 à 77 et les variables 49-50.

Parmi les personnes préférant :		Pourcentage des personnes ayant donné aux allocations familiales (variables 74 à 77)			
		la priorité	la 2è place	la 3è place	la 4è place
"une augmentation des allocations en espèces"(49)	H	66%	17%	9%	8%
	F	68%	18%	8%	6%
"une augmentation des prestations en nature" (50)	H	25%	17%	24%	35%
	F	30%	17%	21%	32%

entre les variables 78 à 81 et les variables 49-50

Parmi les personnes préférant :		Pourcentage des personnes ayant donné aux prêts pour le logement (variables 78 à 81)			
		la priorité	la 2è place	la 3è place	la 4è place
"une augmentation des allocations en espèces"(49)	H	23%	45%	17%	15%
	F	18%	49%	19%	14%
"une augmentation des prestations en nature" (50)	H	41%	20%	20%	19%
	F	30%	29%	24%	17%

Il y a indépendance - plus encore en ce qui concerne les femmes - entre ces attitudes et les attitudes concernant l'adoption éventuelle de critère de ressources pour l'attribution de certaines prestations. Ces attitudes s'échelonnent le long du second axe d'inertie. Du côté positif l'opposition au plafonnement du salaire unique (2) ou (5), ou la fixation d'un plafond très élevé, supérieur à 5000 francs par mois (9), le maintien du système actuel des allocations familiales (42), à un moindre titre l'attribution d'allocations-vacances à tous les enfants (82). Les attitudes opposées à chacune de ces positions se retrouvent au bas des graphiques : plafonnement des allocations familiales (43) ou calcul en fonction inverse des ressources (44), ou plafond très bas pour le salaire unique (3), (6) et (7), allocations vacances à quelques uns seulement (83).

Chez les femmes l'opposition à un critère de ressources est associée à deux types d'attitudes qui en précisent la signification :

a) option pour le financement privé (et non financement de l'Etat) pour la construction de logements, l'aménagement des routes, les transports en commun (54 - 58 - 62).

b) négation de l'inégalité des chances face à l'instruction, du rôle de la situation des parents dans celle des enfants (67 - 72 - 73).

Chez les hommes l'association avec les réponses a) subsiste mais un peu moins marquée, celle avec les réponses b) n'apparaît pas. En outre il n'y a pas tout à fait indépendance entre l'opposition à un critère de ressources et l'option pour les allocations en espèces.

I-3. Introduction de variables exogènes.

A quelles variables socio-économiques sont liées les attitudes précédentes?de quels groupes sociaux émanent-elles ?

- Les partisans d'une augmentation des allocations en espèces (à gauche du premier axe), attachés surtout aux allocations familiales et à l'allocation de salaire unique sont les familles les plus pauvres : manoeuvres, ouvriers spécialisés, les familles dans lesquelles la mère reste au foyer et les familles résidant dans les villes de province, petites et moyennes villes. Pour elles les questions de vacances passent au second plan, leurs problèmes sont des problèmes de vie quotidienne : elles attachent plus d'importance à l'allocation de logement qu'à des prêts d'accession. C'est dans ce seul groupe que l'on trouve un intérêt pour les prêts d'équipement ménager.

- Les partisans d'une augmentation des prestations en nature (qui donnent priorité au problème du logement - surtout les hommes - aux prêts d'accession, aux vacances) sont à l'inverse les familles les plus aisées, surtout les habitants des grandes villes, les parisiens et les familles dont la mère exerce actuellement une activité professionnelle. C'est en particulier une attitude fréquente chez les cadres (femmes) et les employées.

Cette opposition reflète tout à la fois les disparités de niveau de vie et de condition de vie. Chez les plus défavorisés les inégalités économiques sont certainement plus durement ressenties lorsque la femme ne travaille pas - on a vu les différences de ressources que cela induisait. Dans ce cas équipements, tarifs préférentiels... semblent du superflu. Le premier problème est celui des ressources. Pour d'autres familles ce sont les contraintes liées aux conditions d'urbanisation : exigüité des logements, promiscuité, temps de trajets... mais il n'est pas douteux que la sensibilité à ces problèmes implique déjà un niveau de vie minimum. Les tenants des prestations en nature attachent plus d'importance que les autres aux domaines de consommation les plus élastiques par rapport au revenu.

Les variations le long du premier axe en fonction du niveau de vie, sont moins sensibles dans les réponses des hommes que dans celles des femmes. Chez les premiers les cadres supérieurs se situent nettement dans la partie supérieure du graphique, vers l'opposition à la sélectivité, au

lieu de se situer vers la droite. L'opposition des hommes à l'institution d'un critère de ressources plus fréquente aux statuts supérieurs, s'explique probablement par la crainte de perdre des avantages acquis (l'opposition à la formule du plafonnement est moins rare que l'opposition à un calcul en fonction des ressources). Tandis que chez les femmes l'attitude hostile à la sélectivité tiendrait plutôt à une attitude individualiste, assez peu corrélée avec les variables socio-économiques classiques.

I-4. Essai d'interprétation des non-réponses

Les défauts de représentativité des enquêtes ne tiennent pas principalement aux erreurs d'échantillonnage, mais à des causes moins maitrisables : une certaine sélection s'opère par le biais des refus (1) d'une part et d'autre part les réponses ne sont pas indépendantes des circonstances sociales de l'interview.

L'analyse des non-réponses ou des refus partiels permet d'améliorer un peu notre connaissance sur la nature de ces distorsions. Voici très brièvement quelques exemples.

Les techniques d'analyse des données facilitent l'interprétation de ces non-réponses à condition évidemment qu'elles ne soient pas trop nombreuses. Le graphique 3 ci-après déjà commenté a été obtenu à partir d'un traitement des seules réponses exprimées - les autres ayant été réparties aléatoirement - La projection de ces non-réponses en variables illustratives est alors très suggestive : l'ensemble des non-réponses aux ques-

---

(1) Il n'est évidemment pas facile de caractériser les refus proprement dits. On dispose sur eux d'informations très sommaires. Relevons cependant que le taux de refus est en raison inverse des liens économiques avec les caisses d'allocations familiales : 36% chez les familles allocataires, 46% pour les couples touchant des allocations prénatales, 60% ou davantage pour les autres personnes enquêtées (jeunes ménages ou célibataires). Ajoutons qu'à l'intérieur de l'ensemble des familles allocataires, le taux de refus diminue lorsque le nombre d'enfants augmente - alors que le caractère astreignant des interviews et la perte de temps qu'ils représentent pourraient éprouver davantage les mères de familles nombreuses - il est plus faible parmi les familles percevant l'allocation logement. L'échantillon représente au total un ensemble de personnes plus intégrées au système des prestations familiales que la population-mère.

tions touchant à la sélectivité se projette dans la partie supérieure du graphique du côté des opposants à l'institution d'un critère de ressources (1).

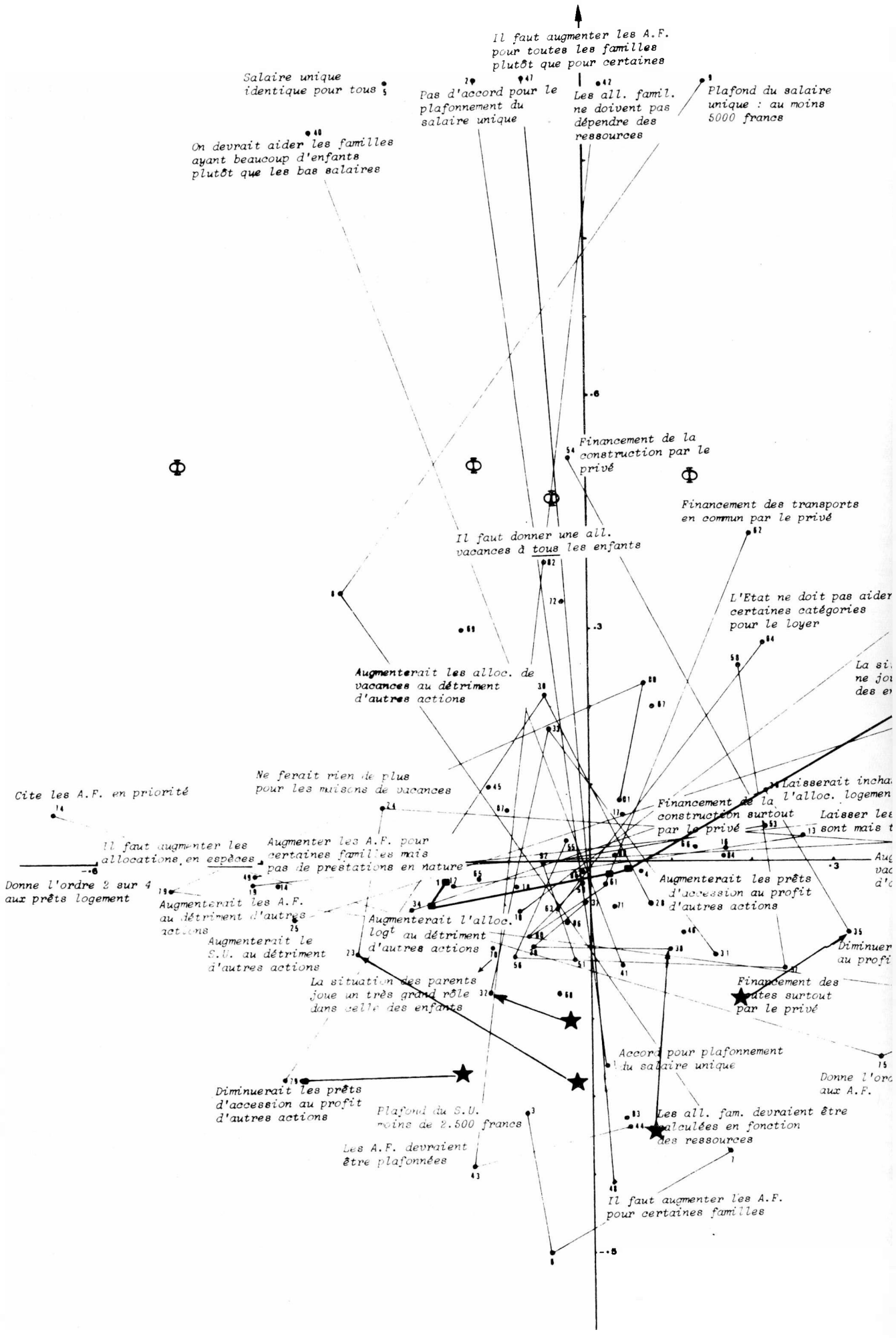
Un autre ensemble de non-réponses est intéressant à analyser. Celui qui concerne la question : s'il fallait diminuer certaines actions des CAF au profit d'autres, lesquelles augmenteriez-vous, lesquelles diminueriez-vous... Suit une énumération de 7 types d'actions en faveur des familles. Comme on s'y attendait il y a eu plus d'"augmentation" que de "diminution", les personnes interrogées répugnant à diminuer quoi que ce soit. Tout au plus on "laisserait inchangé". On observe sur le graphique 4 la position de la non-réponse à chacune des rubriques, toujours plus proche (et du même côté par rapport à F2) du point "diminuerait" (noté par une flèche) que du point "augmenterait". Autrement dit la non-réponse traduit bien une hésitation plus grande à "diminuer" quelque chose et peut être même une crainte des conséquences possibles, la réponse "augmenterait" arrivant plus spontanément.

---

(1) L'analyse de la signification des non-réponses est en cours dans le cadre de la convention de recherche CORDES n° 40/1972 : attitudes par rapport au travail des femmes. Recherches critiques à partir des questionnaires de l'enquête CNAF (1971).

Il est trop tôt pour donner des résultats systématiques, mais l'exemple ci-dessus, d'une non-réponse manifestant une opposition qu'on n'ose exprimer, se reproduit fréquemment.

Voici un autre exemple simple : à la question "Trouvez-vous souhaitable que des adolescents fassent partie d'associations de lycéens ?" 24,5% des femmes et 17,8% des hommes ne répondent pas. Si on restreint l'échantillon aux seules personnes ayant répondu "non" à la question "Encourageriez-vous vos enfants à participer à des groupements ?" (question posée plus tard) les non-réponses précédentes deviennent respectivement 31,5% et 24,2%.


Il faut augmenter les A.F. pour toutes les familles plutôt que pour certaines
Salaire unique identique pour tous
Pas d'accord pour le plafonnement du salaire unique
Les all. famil. ne doivent pas dépendre des ressources
Plafond du salaire unique : au moins 5000 francs
On devrait aider les familles ayant beaucoup d'enfants plutôt que les bas salaires
Financement de la construction par le privé
Financement des transports en commun par le privé
Il faut donner une all. vacances à tous les enfants
L'Etat ne doit pas aider certaines catégories pour le loyer
Augmenterait les alloc. de vacances au détriment d'autres actions
Ne ferait rien de plus pour les maisons de vacances
Cite les A.F. en priorité
Financement de la construction surtout par le privé
Il faut augmenter les allocations en espèces
Augmenter les A.F. pour certaines familles mais pas de prestations en nature
Donne l'ordre 2 sur 4 aux prêts logement
Augmenterait les prêts d'accession au profit d'autres actions
Augmenterait les A.F. au détriment d'autres actions
Augmenterait l'alloc. logt au détriment d'autres actions
Augmenterait le S.U. au détriment d'autres actions
La situation des parents joue un très grand rôle dans celle des enfants
Financement des ...tes surtout par le privé
Accord pour plafonnement du salaire unique
Diminuerait les prêts d'accession au profit d'autres actions
Plafond du S.U. moins de 2.500 francs
Les all. fam. devraient être calculées en fonction des ressources
Les A.F. devraient être plafonnées
Il faut augmenter les A.F. pour certaines familles

Graphique 3

ATTITUDES DES FEMMES PAR RAPPORT
AUX PRESTATIONS FAMILIALES

1383 personnes - 88 variables - 29 questions

en italique : les 88 variables analysées

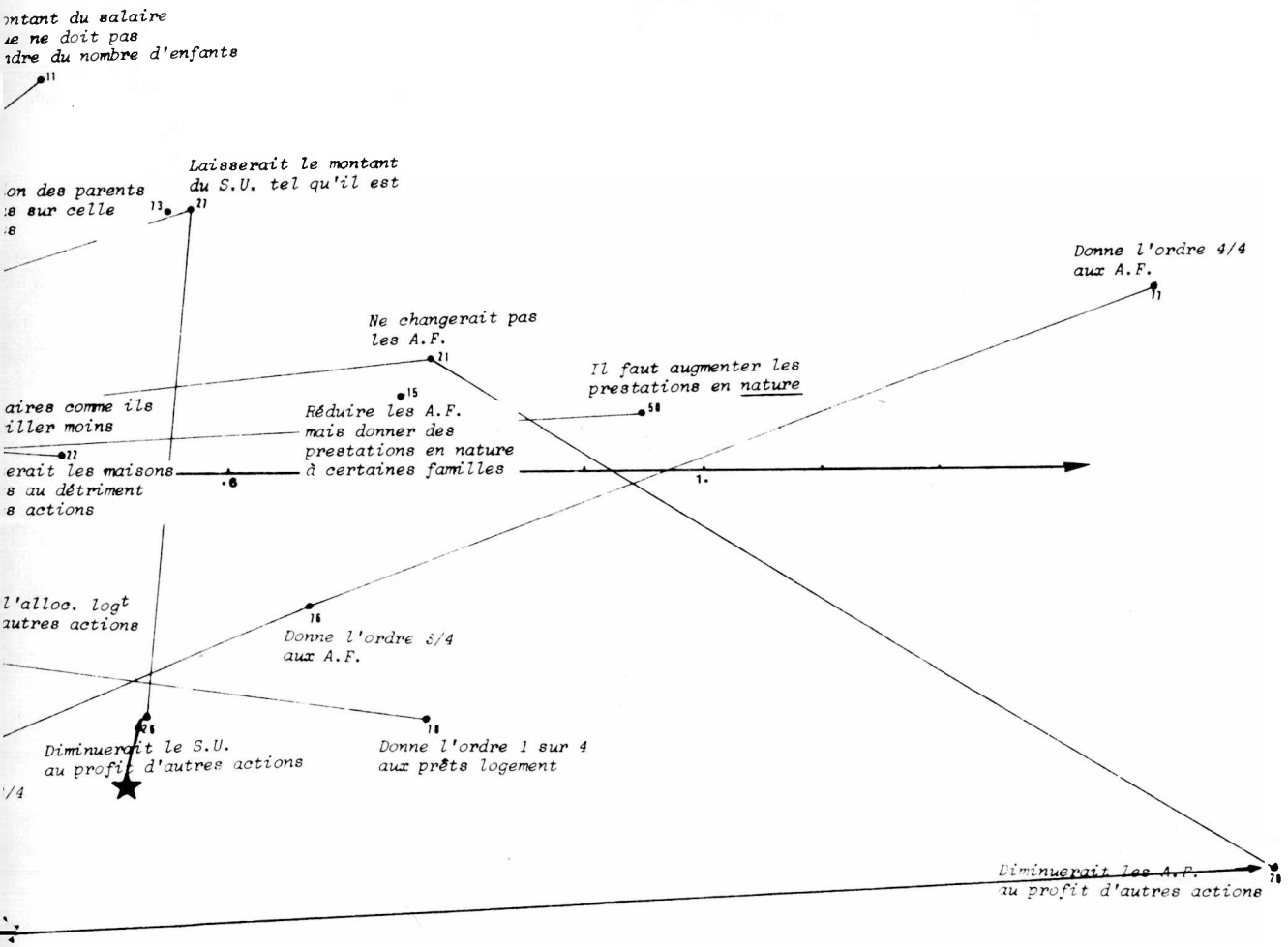
en caractère droit : quelques variables illustratives

en trait fort : le revenu, croissant de 1 à 10

Projection des non-réponses en variables illustratives

Φ non-réponses à 4 questions sur la sélectivité des prestations

★ non-réponses à 7 questions proposant un choix entre augmentation ou diminution



ANNEXE 1

ATTITUDES PAR RAPPORT AUX PRESTATIONS FAMILIALES

Liste des 88 questions analysées

		Nombre de réponses*
A propos de la suppression du salaire unique pour les familles ayant des salaires élevés. Etes-vous d'accord sur cette nouvelle mesure :		
1	oui	1312
2	non	211
		(22)
Auriez-vous choisi une limite différente :		
3	- inférieure à 2.500 francs	225
4	- non, même limite	638
5	- pas de limite, la même chose pour tous	121
6	- 2.500 à 3.000 francs	272
7	- 3.000 à 4.000 francs	141
8	- 4.000 à 5.000 francs	13
9	- 5.000 francs ou davantage	14
		(121)
Le montant de l'allocation de salaire unique devrait-il augmenter à votre avis, selon le nombre des enfants :		
10 -	oui	1286
11 -	non	246
		(13)
Même si on ne s'en aperçoit pas toujours, c'est un fait que le niveau de vie augmente, que certains secteurs, surtout dans l'industrie augmentent leur production grâce au progrès technique. Pour répartir les bénéfices quelle serait à votre avis la meilleure solution :		
12 -	augmenter les salaires et que les gens travaillent autant	943
13 -	laisser les salaires comme ils sont mais que les gens travaillent moins	568
		(34)
Quelle solution trouvez-vous préférable :		
14 -	augmenter les allocations pour certaines catégories de familles mais ne pratiquer aucune réduction de tarif pour les services tels que crèches, garderies, gardes et activités du jeudi, colonies de vacances...	1014
15 -	réduire pour tout le monde les allocations mais faciliter l'utilisation des services à certaines catégories de familles (par des bons, prestations spéciales, etc...)	487
		(44)

* *Pour 1545 femmes, entre parenthèses le nombre des non-réponses. Ces non-réponses ne sont pas incluses comme telles dans l'analyse mais réparties de façon aléatoire entre les réponses possibles.*

L'analyse décrite précédemment porte sur 1383 couples, parmi les 1545 ; la différence concerne les couples dont l'homme n'a pas répondu au questionnaire.

		Nombre de réponse
Si l'on diminuait les allocations familiales, mais qu'on facilite par des prestations en nature (bons...) l'utilisation des services et équipements pour enfants (crèches, garderies, activités du jeudi garderies aérées, équipements de vacances...). Pensez-vous que cela entraînerait un changement du nombre des naissances chez les familles moins aisées :		
16 -	une diminution	419
17 -	une augmentation	191
18 -	pas de changement	874
		(61)
Les caisses d'allocations familiales aident les familles de différentes façons. On énumère ci-dessous les principales formes d'aide. S'il fallait en diminuer certaines au profit d'autres, lesquelles augmenteriez-vous, lesquelles diminueriez-vous ?		
	- Les allocations familiales :	
19 -	vous augmenteriez	1033
20 -	vous diminueriez	41
21 -	vous laisseriez inchangé	450
		(21)
	- La construction de maisons de vacances :	
22 -	vous augmenteriez	537
23 -	vous diminueriez	369
24 -	vous laisseriez inchangé	582
		(57)
	- L'allocation de salaire unique :	
25 -	vous augmenteriez	937
26 -	vous diminueriez	147
27 -	vous laisseriez inchangé	425
		(36)
	- Les prêts aux familles pour l'achat de logement :	
28 -	vous augmenteriez	946
29 -	vous diminueriez	134
30 -	vous laisseriez inchangé	417
		(48)
	- Les allocations spéciales de vacances :	
31 -	vous augmenteriez	472
32 -	vous diminueriez	329
33 -	vous laisseriez inchangé	684
		(60)
	- L'allocation de logement :	
34 -	vous augmenteriez	789
35 -	vous diminueriez	93
36 -	vous laisseriez inchangé	607
		(56)
	- Les prêts aux familles pour l'achat d'équipement ménager :	
37 -	vous augmenteriez	371
38 -	vous diminueriez	447
39 -	vous laisseriez inchangé	671
		(56)

	Nombre de réponses
S'il y avait une augmentation des sommes destinées aux familles devrait-elle être attribuée :	
40 - plutôt aux familles ayant beaucoup d'enfants quel que soit le niveau des ressources	172
41 - plutôt aux familles ayant de faibles ressources quel que soit le nombre d'enfants	1346
	(27)
Actuellement les allocations familiales sont les mêmes pour toutes les familles quel que soit leur revenu. Trouvez-vous :	
42 - que c'est bien ainsi	328
43 - qu'elles devraient être supprimées à partir d'un certain salaire	317
44 - qu'elles devraient être calculées en fonction du salaire (diminuées progressivement lorsque le salaire augmente)	889
	(11)
A votre avis dans lequel de ces deux domaines l'Etat devrait-il faire porter davantage son effort :	
45 - plutôt sur la construction et l'entretien des routes, parking..	685
46 - plutôt sur le développement des transports en commun	803
	(57)
Quelle serait à votre avis la meilleure utilisation d'une augmentation éventuelle des fonds destinés aux familles :	
47 - augmenter les allocations familiales pour toutes les familles, mais faiblement.	439
48 - augmenter davantage les allocations familiales mais pour certaines catégories de familles seulement	1078
	(28)
Et des deux solutions suivantes, laquelle vous semble préférable	
49 - une augmentation des allocations en espèces	1061
50 - une attribution de prestations en nature pour l'utilisation de certains services ou équipements (gardes, vacances)	469
	(15)
La collectivité (institutions diverses, Caisses d'Allocations) ne peut pas tout financer ou subventionner. Dans certains domaines le financement est privé ou il y a partage. Pour les fonctions énumérées ci-dessous, comment feriez-vous le partage :	
- La construction de logements :	
51 - financement collectif seulement	481
52 - surtout collectif mais avec partage	645
53 - surtout privé mais avec partage	245
54 - financement privé seulement	84
	(90)
- L'aménagement des routes :	
55 - financement collectif seulement	838
56 - surtout collectif mais avec partage	344
57 - surtout privé mais avec partage	137
58 - financement privé seulement	124
	(102)

		Nombre de réponses
	- Les transports en commun :	
59	- financement collectif seulement	711
60	- surtout collectif mais avec partage	399
61	- surtout privé mais avec partage	183
62	- financement privé seulement	146
		(106)
	Estimez-vous normal que l'Etat prenne en charge en partie ou en totalité le loyer ou les charges de certaines catégories de familles ou de personnes	
63	- oui	1257
64	- non	265
		(23)
	Laquelle des solutions suivantes vous paraît la meilleure pour améliorer les conditions de logement :	
65	- que l'Etat participe lui-même à la construction de logements	721
66	- que les initiatives privées de construction soient encouragées	738
		(86)
	Pensez-vous qu'à l'heure actuelle tous les enfants ont les mêmes possibilités de faire des études poussées, qu'ils soient enfants d'ouvriers ou de cadres ?	
67	- oui	616
68	- non	858
69	- ne se prononce pas	66
		(5)
	Pensez-vous que la situation des parents, leurs relations ou leur savoir-faire jouent un très grand rôle dans la situation des enfants	
70	- c'est très important	510
71	- c'est assez important	718
72	- cela joue peu	219
73	- cela n'a aucune importance	81
		(17)
	S'il y avait une augmentation des fonds destinés aux familles à quoi devrait-elle être consacrée selon vous, parmi les quatre possibilités suivantes	
	a - à une augmentation des allocations familiales	
	b - à la création d'équipements pour enfants en bas âge	
	c - à la création d'équipements de vacances et de loisirs	
	d - à des prêts aux familles pour le logement, l'équipement ménager	
	On retient les deux rubriques suivantes :	
	- A une augmentation des allocations familiales	
74	- ordre de préférence 1	828
75	- ordre de préférence 2	276
76	- ordre de préférence 3	194
77	- ordre de préférence 4	210
		(37)
	- A des prêts aux familles pour le logement, l'équipement ménager	
78	- ordre de préférence 1	352
79	- ordre de préférence 2	597
80	- ordre de préférence 3	330
81	- ordre de préférence 4	224
		(42)

	Nombre de réponses
Laquelle de ces deux solutions jugez-vous préférable :	
82 - donner une allocation vacances modeste mais à tous les enfants	706
83 - donner une allocation plus importante mais à certaines catégories d'enfants seulement	815
	(24)
Près de chez vous, dans votre localité, dans votre quartier... là où vous utilisez ou utiliseriez les services et les équipements correspondants, quel est votre sentiment sur ce que l'on fait pour la famille en général :	
84 - beaucoup de choses	209
85 - pas assez de choses	341
86 - vraiment peu de choses	323
87 - rien du tout	311
88 - ne se prononce pas	351
	(10)

EXEMPLE 2*

II - LES BUDGETS FAMILIAUX DANS LES REGIONS DE LA C.E.E.

Cette analyse est un exemple de traitement classique d'un tableau que l'on peut assimiler à un tableau de contingence. Elle est surtout intéressante par ses résultats :

Le degré de ressemblance entre les profils de consommation des différentes régions est fonction de leur proximité géographique.

En d'autres termes, lorsqu'on traite par l'analyse des correspondances le tableau des structures de budget de tous les groupes socio-géographiques étudiés, on obtient dans le plan des deux premiers facteurs la carte géographique de la C.E.E. - un peu stylisée mais facile à reconnaître : le premier facteur reproduit approximativement l'axe Sud-Nord, Sicile - Hambourg, le second, l'axe Ouest-Est, le troisième axe oppose pour chaque pays, pris séparément, les ouvriers aux employés et fonctionnaires.

A l'intérieur des trois grands pays ; Allemagne, Italie, France, les différences entre ouvriers et employés-fonctionnaires, l'emporte sur les différences régionales, mais surtout elle l'emporte dans certains cas sur la nationalité : les ouvriers néerlandais sont plus proches des ouvriers allemands que ne le sont les employés-fonctionnaires allemands eux-mêmes.

II-1. Présentation des variables analysées.

Une enquête sur les budgets familiaux a été effectuée en 1963 dans les six pays de la communauté. Les résultats en ont été publiés par l'Office Statistique des Communautés Européennes (1). C'est une partie de ces résultats, la publication régionale, que nous analysons ici.

---

(1) Cette enquête a fait l'objet de sept publications : Office Statistique des Communautés Européennes, O.S.C.E. : Budgets Familiaux 1963-1964. Série Spéciale. Les numéros 1 à 6 concernent chacun des six pays séparément. Le volume 7 donne les résultats au niveau des régions pour les trois pays Allemagné, France et Italie.

* Cet exemple est publié en partie, in : Consommation n°1, 1972 p.79-90 et intégralement in : BENZECRI J.P. - L'analyse des données - Tome 2 - DUNOD - 1973.

L'enquête auprès des ménages est la méthode de recueil de données la plus directe et la nomenclature des biens et services de consommation retenue par les experts est suffisamment détaillée et précise pour que chaque rubrique corresponde au même contenu pour chacun des pays étudiés. C'est sans doute un matériau statistique relativement fiable pour des comparaisons internationales. Peu de problèmes de concepts, et si les techniques d'enquêtes ont été quelquefois différentes, il n'y a pas de raison de penser que l'ordre de grandeur des erreurs de mesure soit très différent.

On a ainsi tenté, par curiosité, une comparaison des profils de consommation des ménages de la communauté, comme si celle-ci était un ensemble de 31 régions. Les dépenses moyennes de consommation par ménage figuraient en monnaie nationale dans les publications de l'O.S.C.E. Elles ont été converties ici en unité de comptes commune, le franc belge, au taux de parité monétaire de 1963 (1). Avec la méthode statistique adoptée, l'analyse des correspondances, ces taux ne jouent que comme coefficients de pondération de chacun des pays. Leur modification ne changerait rien aux résultats qui vont suivre.

L'ensemble des salariés (non agricoles) de la Communauté a été divisé en sous-groupes distinguant les ouvriers des employés-fonctionnaires et isolant les grandes régions pour l'Allemagne, la France et l'Italie.

Au total on obtient 55 sous-groupes combinant le statut professionnel et la région. La définition de ces unités socio-géographiques est donnée à l'annexe 1. Le budget des ménages constituant ces unités est décomposé en 65 postes de consommation auxquels on a ajouté le montant des cotisations de Sécurité Sociale et les impôts, soit 67 rubriques dont la liste et le montant sont donnés à l'annexe 2.

Ce travail est l'analyse du tableau dont l'élément x_{ij} est la dépense moyenne annuelle évaluée en francs belges des ménages appartenant à l'unité socio-géographique i pour le poste de consommation j. Ce tableau, trop volumineux (55 lignes , 67 colonnes), est épargné au lecteur.

---

(1) Les taux de conversion sont les suivants :
1 Deutschmark = 12,5 FB - 100 Lires = 8,0 FB - 1 Francs français = 10,2 FB - 1 Florin = 13,8 FB - 1 Franc luxembourgeois = 1 FB.

La méthode utilisée est l'analyse des correspondances.

L'économiste est familiarisé avec la notion de coefficient budgétaire et l'expression suivante retenue comme distance entre deux pays i et i' semblera assez naturelle.

$$d^2(i,i') = \sum_j \frac{x_{..}}{x_{.j}} \left[\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}} \right]^2$$

$-x_{i.}$ étant la consommation totale du pays i, $x_{ij}/x_{i.}$ son coefficient budgétaire pour le poste de consommation j

$-\frac{x_{.j}}{x_{..}}$ est le poids du poste j dans la consommation totale de tous les pays étudiés. On pondère les différences ci-dessus par l'inverse de ce terme, ce qui réduit le rôle éventuellement joué dans d^2 par des postes de consommation disproportionnés par rapport aux autres.

Notre intention n'est pas de remplacer le tableau de consommations par un autre, celui des distances deux à deux, même s'il est un peu plus petit, mais de représenter sur une figure tous les pays disposés de façon à respecter au mieux leurs distances relatives. Ainsi on jugera d'un seul coup d'oeil les pays proches, c'est-à-dire ceux qui auront des profils budgétaires voisins, et les pays éloignés. C'est un des résultats de l'analyse des correspondances, le graphique 2 le résume.

De la même façon on calcule des distances entre les dépenses de consommation. La formule est la même que ci-dessus en intervertissant les indices i et j : il ne s'agira plus de coefficients budgétaires mais de "la part consommée par un pays par rapport à tous les autres pour un produit donné". Deux consommations seront voisines si chaque pays consomme la même proportion de chacune d'elles.

On pourrait se limiter là, sans chercher à savoir pourquoi deux pays sont voisins, c'est-à-dire quelles sont les consommations qu'ils privilégient, quelles sont celles qu'ils rejettent. Pourquoi deux produits sont voisins c'est-à-dire quels sont les pays qui en prennent la plus grande part... ou la plus petite.

La représentation simultanée des deux ensembles de distances est possible. C'est cette représentation que nous commenterons maintenant.

Le graphique 1 est le résultat, limité au plan des deux premiers facteurs, d'une analyse des correspondances effectuée sur le tableau initial. Il contient 63,7% (41,3% et 22,4% pour le premier et le second facteur respectivement) de l'information contenue dans ce tableau (1). C'est la présentation plane qui respecte le mieux les similitudes entre régions d'une part : les régions les plus proches ont des profils de consommation les plus semblables, et entre dépenses d'autre part : les consommations : les consommations les plus proches sont celles qui ont les mêmes amateurs.

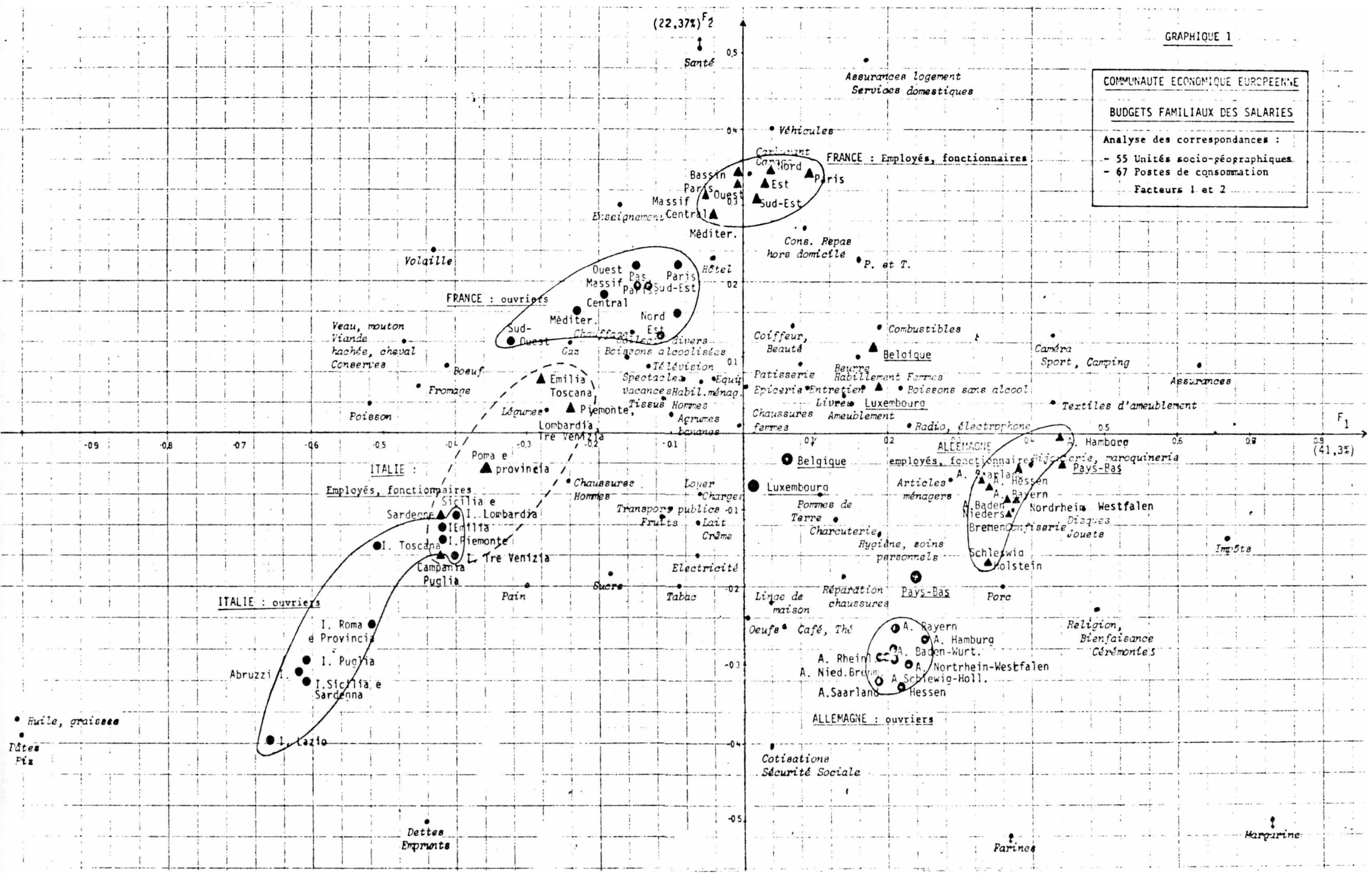
Par contre la distance terme à terme entre une dépense et un pays n'a pas de sens ; mais la position d'un pays par rapport à l'ensemble de toutes les consommations en a un : il se rapproche des postes de budget pour lesquels il a, par rapport aux autres pays, les coefficients budgétaires maxima. L'Italie est ainsi proche des consommations d'huile, riz, pâtes, loin de la margarine ou de la confiserie dont les amateurs sont allemands ou néerlandais...

II-2. Interprétation des axes d'inertie

Vu la disparité des niveaux de consommation totale entre les groupes étudiés (de 1 à 3 entre les ouvriers des Abruzzes et les employés-fonctionnaires parisiens) on aurait pu s'attendre à ce que l'un des axes principaux s'interprète en fonction du niveau de vie :

- sur le premier axe factoriel par exemple, les dépenses semblent s'échelonner en fonction de leur élasticité par rapport au revenu : plus de dépenses alimentaires à gauche, dépenses de loisir ou d'aménagement d'intérieur, ou d'assurances... à droite. (A quelques consommations près que l'on pourrait considérer comme des particularités propres à certains pays expliquant les oppositions : boissons alcoolisées à boissons non alcoolisées, boeuf - fromage - poissons à porc - charcuterie, huile à margarine, légumes à pommes de terre, etc...).

- Les 3ème, 4ème, 5ème et 6ème facteurs représentent respectivement 11,1%, 4,6%, 3,4%, 2,5% de la dispersion totale. Ainsi on peut, dans un espace à 6 dimensions, représenter 85% de l'information contenue dans l'espace initial à 55 dimensions dont les points seraient les dépenses, ou l'espace à 67 dimensions dont les points seraient les pays.


GRAPHIQUE 1
COMMUNAUTE ECONOMIQUE EUROPEENNE
BUDGETS FAMILIAUX DES SALARIES
Analyse des correspondances :
- 55 Unités socio-géographiques
- 67 Postes de consommation
Facteurs 1 et 2
(22,37%) F2
F1
(41,3%)
FRANCE : Employés, fonctionnaires
FRANCE : ouvriers
ITALIE : Employés, fonctionnaires
ITALIE : ouvriers
ALLEMAGNE : employés, fonctionnaires
ALLEMAGNE : ouvriers
Santé
Assurances logement
Services domestiques
Véhicules
Bassin Parisien Ouest
Nord
Est
Paris
Sud-Est
Massif Central
Méditer.
Enseignement
Cons. Repas hors domicile
P. et T.
Volaille
Hôtel
Ouest
Bas Massif Central
Paris Sud-Est
Nord Est
Méditer.
Sud-Ouest
Veau, mouton
Viande hachée, cheval
Conserves
Boeuf
Fromage
Poisson
Chauffage
Gaz
Boissons alcoolisées
Télévision
Spectacles vacances
Habil.ménag.
Tissus Hommes
Agrumes
Bananes
Légumes
Emilia
Toscana
Piemonte
Lombardia
Tre Venizia
Roma e provincia
Sicilia e Sardegna
I. Lombardia
I. Emilia
I. Toscana
I. Piemonte
I. Tre Venizia
Campania Puglia
I. Roma e Provincia
I. Puglia
Abruzzi I.
I. Sicilia e Sardegna
I. Lazio
Huile, graisses
Fêtes Fix
Dettes Emprunts
Pain
Sucre
Tabac
Chaussures Hommes
Loyer
Charges
Transports publics
Fruits
Lait Crème
Electricité
Coiffeur, Beauté
Combustibles
Belgique
Patisserie
Epicerie
Beurre
Habillement Femmes
Entretien
Poissons sans alcool
Livres
Luxembourg
Ameublement
Chaussures femmes
Radio, électrophone
Caméra Sport, Camping
Assurances
Textiles d'ameublement
Articles ménagers
Pommes de Terre
Charcuterie
Hygiène, soins personnels
Linge de maison
Réparation chaussures
Oeufs
Café, Thé
Pays-Bas
Porc
Pâtisserie, maroquinerie
A. Hamburg
A. Saarland
A. Hessen
A. Bayern
A. Baden-Württ.
Nordrhein Westfalen
Bremen
Confiserie
Disques Jouets
Schleswig Holstein
Impôts
Religion, Bienfaisance Cérémonies
A. Rheinl.-Pf.
A. Nied.Brem.
A. Nordrhein-Westf.
A. Schlewig-Holl.
A.Saarland
Hessen
Cotisations Sécurité Sociale
Farines
Margarine

- la direction de la première bissectrice pourrait elle aussi faire croire à des disparités de niveau de vie, du fait du décalage vers le haut des employés fonctionnaires par rapport aux ouvriers. La consommation totale est un estimateur de niveau de vie moins contestable pour des comparaisons faites à l'intérieur d'un pays qu'entre pays. Or cette consommation totale n'augmente pas le long de cet axe. En particulier Rome, où le niveau de consommation est le plus élevé, occupe toujours le centre de l'Italie, sa position géographique (1).

- le troisième axe qui oppose les ouvriers aux employés-fonctionnaires n'est pas non plus un axe de niveau de vie malgré la séparation entre dépenses alimentaires et non alimentaires : les pays se situent approximativement au même niveau sur cet axe pour chacune des strates. Tout se passe comme si, ayant analysé deux groupes sociaux, on obtenait deux analyses identiques, les deux premiers facteurs, avec un décalage, le troisième facteur.

L'essentiel de l'analyse est contenu dans le premier plan et son interprétation est fort simple : l'analogie avec la carte géographique du marché commun est frappante. Elle l'est davantage sur le graphique 2 où le premier facteur, axe Nord-Sud, est présenté verticalement.

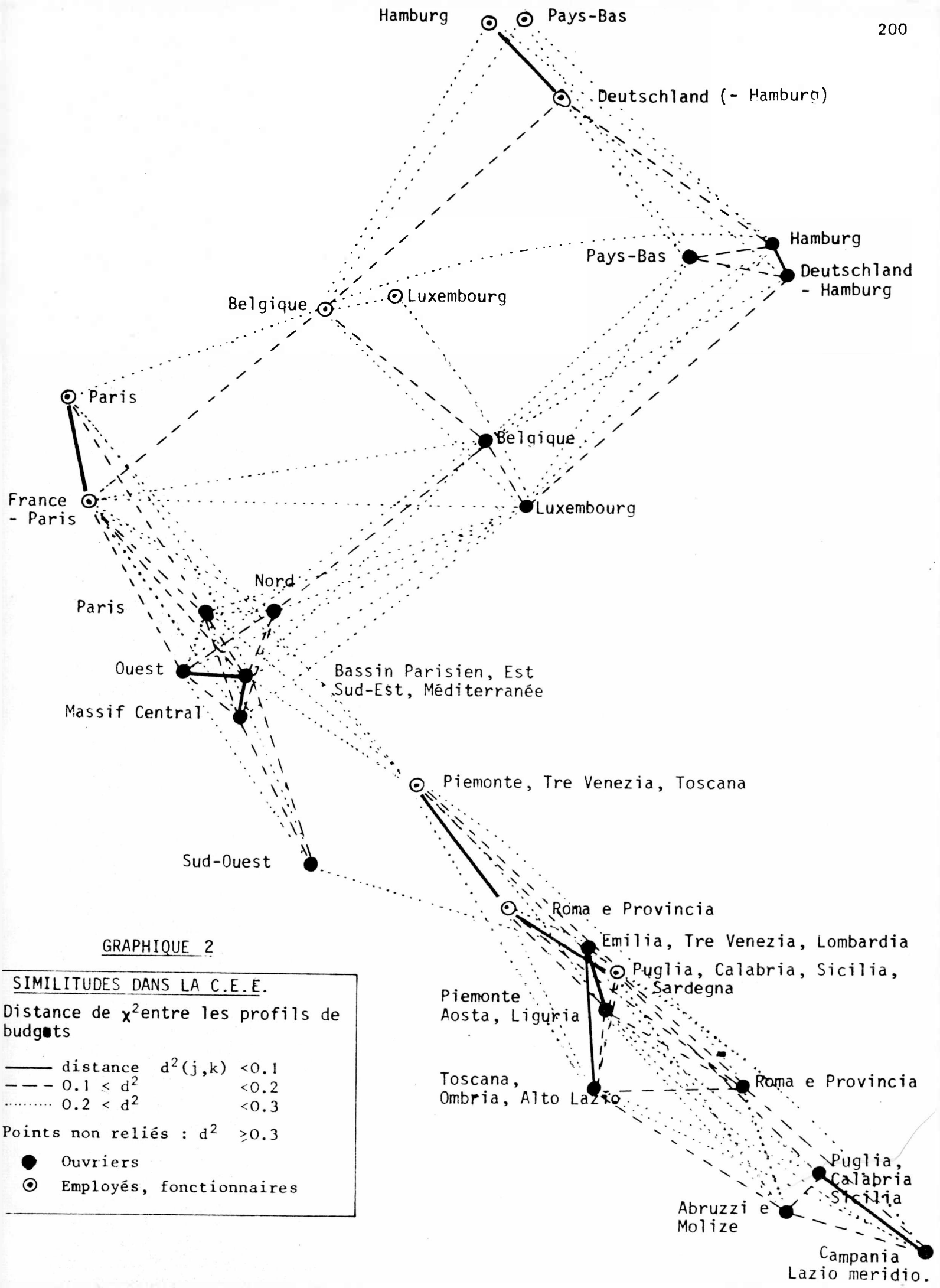
II-3. Proximité entre dépenses de consommation.

Il ne se forme pas de partition de dépenses à proprement parler, telleque chaque classe serait caractéristique d'un pays ou d'une région donnée (2). Mais plutôt une évolution continue avec quelques rubriques excentriques les impôts par exemple (Pays-Bas - Allemagne) ou les cotisations de Sécurité Sociale qu'on notera opposées aux dépenses de santé (France).

---

(1) On a vérifié la stabilité du plan des deux premiers facteurs pour chacune des deux catégories sociales analysées séparément, dans les deux cas la configuration est la même. On a également vérifié la stabilité de ce plan lorsqu'on enlève des postes de consommation très typiques tels la margarine, l'huile, les pâtes....

(2) Avec un seuil de distance très faible ($d^2 \leqslant 0.06$) on trouve les achats de véhicules et les dépenses de carburants propres à la France et quelques autres associations qui se regroupent toutes à un seuil légèrement plus élevé.



GRAPHIQUE 2

SIMILITUDES DANS LA C.E.E.

Distance de χ^2 entre les profils de budgets

———	distance $d^2(j,k)$	<0.1
– – –	$0.1 < d^2$	<0.2
·····	$0.2 < d^2$	<0.3
Points non reliés :	d^2	≥ 0.3

● Ouvriers

⊙ Employés, fonctionnaires

L'axe Nord-Sud oppose les consommations exprimant un intérêt pour le logement, tant du point de vue de l'aménagement et du confort que du point de vue des activités, aux consommations manifestant davantage la vie à l'extérieur.

Ainsi les dépenses suivantes :

- linge de maison	(Allemagne (1) - Italie (1)
- textiles d'ameublement	(Pays-Bas ou Allemagne - Italie)
- articles et accessoires ménagers	(Pays-Bas ou Allemagne - Italie)
- ameublement	(Pays-Bas - Italie)
- entretien	(Luxembourg - Italie)

et également celles de :

- livres	(Pays-Bas - Italie ou France)
- disques, jouets	(Pays-Bas - Italie)
- radio, électrophones	(Pays-Bas ou Allemagne - Italie)

se situent à droite du premier axe (et même sous la première bissectrice), tandis que les dépenses ci-dessous se situent à gauche :

- transports publics	(Italie - Luxembourg ou Pays-Bas)
- hôtel, logement de vacances	(Italie - Luxembourg)
- spectacles, vacances (2)	(France ou Italie - Belgique)

Cette opposition pourrait aussi traduire un niveau de technicité plus élevé dans les pays du Nord.

On remarquera l'opposition entre les dépenses d'habillement pour hommes (vêtements et surtout chaussures), situées à gauche et celles pour femmes à droite. Les premières sont supérieures aux secondes en France et en Belgique et surtout en Italie et la disparité est encore plus grande chez les ouvriers que chez les employés-fonctionnaires. Quant aux différences bien connues des régimes alimentaires, elles sont exprimées de façon presque caricaturale sur le graphique 1.

---

(1) Le premier pays de la parenthèse est celui où le coefficient budgétaire de la dépense concernée est le plus élevé, le second celui où il est le plus bas.

(2) Non compris frais d'hôtel et transports vacances.

II-4. Comparaison des distances entre groupes socio-géographiques

Cette distance est l'expression d^2 calculée pour tous les couples formés à partir des 55 unités socio-géographiques de base. L'importance des deux premiers facteurs est telle que ces distances sont à peu près respectées dans la figure 1. Mais cette figure donnant la représentation simultanée de l'espace pays et de l'espace consommation, contient trop d'informations pour une lecture claire des distances seules ; celles-ci sont reprises au graphiques 2.

Les 55 unités socio-géographiques de départ ont été réduites à 28 par regroupement des unités telles que $d^2 \leqslant 0.09$.

Les plus petites distances intéressent toujours un même pays et un même groupe social. On est donc passé de 55 à 28 groupes en regroupant les ouvriers de certaines régions d'un même pays, ou les employés-fonctionnaires. Ce premier stade concerne donc seulement les trois grands pays.

L'Allemagne est le pays le plus homogène des trois. C'était visible sur le graphique 1. Il ne subsiste que deux régions celle de Hambourg et toutes les autres ($d^2 \leqslant 0.08$).

Pour la France le regroupement est différent selon qu'il s'agit des ouvriers ou des employés-fonctionnaires. Pour les premiers on passe des neuf régions initiales à six en regroupant seulement : Bassin Parisien - Est - Sud-Est - Midi ($d^2 \leqslant 0.008$). Pour les seconds deux régions restent distinctes : Paris et le reste de la France ($d^2 \leqslant 0.09$).

L'Italie est le pays où la diversité des comportements est la plus grande. Des dix régions analysées initialement pour les ouvriers on en conserve sept, regroupant seulement :

Lombardia - Tre Venezie - Emilia	($d^2 \leqslant 0.08$)
Puglia - Sicilia e Sardegna	($d^2 = 0.07$)

Les employés-fonctionnaires se répartissent en trois régions au lieu de cinq.

L'Italie septentrionale : Piemonte, Tre Venezie	($d^2 = 0.06$)
L'Italie médiane : Roma e Provincia	inchangée
L'Italie méridionale : Puglia, Sicilia e Sardegna	($d^2 = 0.08$)

Ce premier stade de regroupement fait, on obtient entre les 28 unités restantes les distances relatives figurant au graphique 2. (L'annexe 4 donne les chiffres correspondants).

La figure 2 ressemble bien à la carte du marché commun avec aux deux extrémités : Nederland - Hamburg et Campania - Abruzzi - Calabria (1)

Les plus petites distances ($d^2 \leq 0.1$, en traits forts) n'existent qu'à l'intérieur d'un pays et d'un groupe social. Le découpage géographique de l'Allemagne pourrait être supprimé complètement, les différences régionales de comportement budgétaire y sont négligeables comparées à ce qui existe en France et surtout en Italie.

Il y a dans les pays du Nord de la Communauté une proximité plus grande entre ouvriers de pays différents qu'entre les ouvriers et les employés-fonctionnaires d'un même pays.

- Les ouvriers néerlandais ont un comportement plus proche des ouvriers allemands (d^2 = 0.17 ou 0.18) que des employés-fonctionnaires de leur pays (d^2 = 0.23). C'est très visible sur le graphique 1 où les constellations "ouvriers" et "employés-fonctionnaires" sont bien distinctes.

- Les ouvriers belges et luxembourgeois proches entre eux, sont relativement éloignés des employés-fonctionnaires de leur propre pays.

Ce n'est plus vrai pour l'Italie et la France, même en prenant les régions extrêmes : les employés-fonctionnaires ne se ressemblent pas plus que ne se ressemblent les ouvriers, alors que la dispersion de la consommation totale serait plutôt plus faible chez les premiers : les distances moyennes entre régions sont approximativement les mêmes pour les deux catégories sociales.

---

(1) La disposition des pays sur le graphique 2 est la même que sur le graphique 1 - a regroupement près indiqué ci-dessus. Par rapport au graphique 1, l'information supplémentaire-est la distance d^2 indiquée par des traits-tirets ou points qui donnent une idée des proximités entre pays dans l'espace complet - non limité aux deux premiers facteurs.

Bref, l'explication des différences de comportement de consommation des ménages de la Communauté semble relever plus de la géographie que de l'économie, - malgré les disparités de niveau de vie. L'Allemagne seule semble très homogène et proche des Pays-Bas.

Nous venons de remplacer un tableau de chiffres par une présentation graphique à la fois très suggestive et réduisant l'information à ses éléments les plus significatifs.

On pourrait imaginer de substituer ces analyses à la plupart des publications de chiffres, les comptes-rendus d'enquêtes en particulier. Beaucoup trop lourdes pour les lecteurs qui en attendent un résultat, elles sont presque toujours insuffisantes comme produit intermédiaire, pour le chercheur qui souhaite les analyser. Dans ce dernier cas, l'échange des fichiers de base, carthotèques ou bandes magnétiques, plus économiques et facteur de progrès, devrait être depuis longtemps généralisé.

Du point de vue de l'utilisateur final, l'effort de passer au crible les données serait certainement payé de retour : l'information ramenée à ses traits les plus pertinents, plus largement diffusée : les chiffres réduits à leur valeur qui n'est que relative ; le statisticien lui-même aurait peut-être des surprises, les critères de tri traditionnels n'étant pas toujours les plus efficaces.

ANNEXE 1

COMMUNAUTE ECONOMIQUE EUROPEENNE

Découpage régional étudié

ANNEXE II

PARTITION GEOGRAPHIQUE ETUDIEE

Nombre de ménages enquêtés ,n, et consommation totale par ménage, CT.

	Ouvriers		Employés, fonctionnaires	
	n	CT[1]	n	CT[1]
ALLEMAGNE				
Schleswig-Holstein	200	1321	132	1659
Hamburg	156	1544	158	2146
Niedersachsen'Bremen	705	1313	391	1725
Nordrhein-Westfalen	1675	1450	831	1885
Hessen	409	1335	283	1821
Rheinland-Pfalz	304	1331	238	1793
Baden-Württemberg	687	1369	427	1874
Bayern	814	1324	476	1764
Saarland	135	1469		
		1385		1839
FRANCE				
Paris	794	1845	985	2582
Bassin parisien	692	1445	392	1908
Nord	586	1490	292	2236
Est	673	1541	370	2117
Ouest	525	1235	293	1793
Massif-Central	194	1274	469	1845
Sud-Ouest	623	1339		
Sud-Est	703	1506	388	2067
Méditerranée	523	1471	421	2147
		1506		2173
ITALIE				
Piemonte, Valle d'Aosta, Liguria	710	1249		
Lombardia	934	1378	671	1879
Tre Venezie	682	1178		
Emilia, Romagna, Marche	389	1142		
Toscana, Umbria, Alto Lazio	407	1162	253	1788
Lazio meridionale, Campania	436	935		
Abruzzi e Molize	220	865	300	1523
Puglia, Basilicata, Calabria	268	974		
Sicilia e Sardegna	790	956	253	1472
Roma e provincia	553	1397	379	1954
		1163		1770
BELGIQUE	2786	1516	1611	2277
PAYS-BAS	2619	1437	1572	2237
LUXEMBOURG	1084	1710	723	2180

1 - Unité : 100 francs belges.

ANNEXE III

NOMENCLATURE DES BIENS ET SERVICES RETENUS ET DEPENSE MOYENNE CORRESPONDANTE[1]

1	Pain	3559	35	Loyer et charges	13573
2	Patisserie (fraîche et autre)	1811	36	Combustibles liquides et solides	3777
3	Riz, pâtes alimentaires	1021	37	Electricité	1917
4	Céréales et farines	592	38	Gaz (de ville, butane, propane...)	1445
5	Boeuf	3757			
6	Porc	1682	39	Réparation du logement, frais de chauffage	1791
7	Charcuterie, lard, jambon abats	3958	40	Ameublement	3511
8	Volailles, lapin, gibier	1813	41	Linge de maison, literie	941
9	Veau, mouton, plats cuisinés	3024	42	Textiles d'ameublement	466
			43	Equipement ménager	2888
10	Poissons, mollusques et crustacés	1498	44	Articles et accessoires ménagers, vaisselle...	1025
11	Fromage	1828	45	Entretien, blanchissage, teinturerie	2376
12	Lait, crème, fromage blanc, bouillies pour bébés	3093	46	Assurances mobilières, services domestiques, location d'appareils...	1400
13	Oeufs	1440			
14	Beurre	2127	47	Articles de toilette, parfumerie, savons...	1075
15	Margarine	657			
16	Huile, graines animales et végétales	1283	48	Salon de coiffure, de beauté	1199
			49	Services médicaux, pharmacie	2670
17	Agrumes et bananes	1192	50	Transports publics	1787
18	Fruits	2035	51	Véhicules	5126
19	Pommes de terre	994	52	Frais d'utilisation de véhicules	5931
20	Légumes	3515			
21	Sucre	952	53	P. et T., déménagement	1567
22	Confiture, confiserie	1416	54	Livres, journaux, périodiques	1369
23	Epices, ingrédients, fécule	608			
			55	Télévision	1330
24	Café, thé	1740	56	Radio, tourne-disque...	585
25	Boissons sans alcool	822	57	Articles sport, camping, photo...	450
26	Boissons alcoolisées, alcool	4111			
			58	Disques, plantes, jouets, animaux	1566
27	Repas et consommations à l'extérieur du domicile	3563	59	Spectacles, vacances[2], services de loisirs	3028
28	Tabac	2839			
29	Vêtements hommes, garçonnets	5757	60	Enseignement	1584
			61	Bijouterie, articles divers	1017
30	Vêtements femmes, fillettes	5528	62	Assurances privées	2658
			63	Religion, cérémonies, dons	2005
31	Tissus, layette, façon et réparation de vêtements	1674	64	Dettes et emprunts	1676
			65	Hotel, logement de vacances	913
32	Chaussures hommes, garçonnets	1414		Consommation totale	145786
33	Chaussures femmes, fillettes	1364	66	Cotisation Sécurité Sociale	10000
34	Réparation de chaussures	473	67	Impôts	7761

1 - En franc belge, par ménage et par an, pour l'ensemble des salariés de la C.E.E.(1963)
2 - Dépenses de vacances exception faite du transport (dans 53) et de l'hôtel (dans 65).

ANNEXE 4

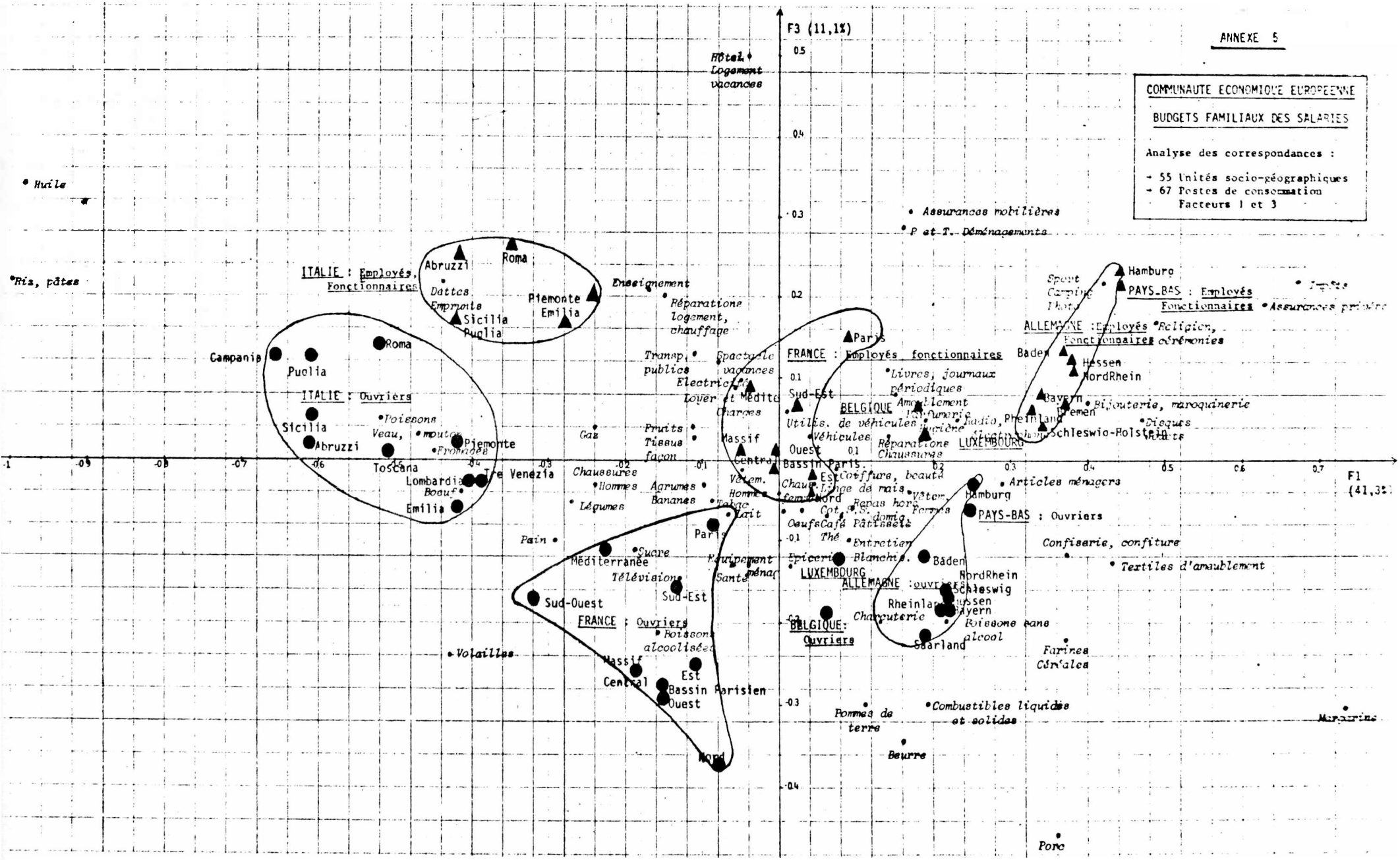
DISTANCE DE χ^2 ENTRE LES UNITES SOCIOGEOGRAPHIQUES[1]

	DESIGNATION DES VARIABLES	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Pays-Bas (Employés, Fonction.)		0•16	0•18	0•22	0•27	0•37	0•28	0•47	0•49	0•49	0•42	0•46	0.64	0.84
2	Hamburg " "	0•16		0•06	0•40	0•25	0•37	0•26	0•39	0•48	0•47	0•34	0•41	0.52	0.73
3	Deutschland (- Hamburg) "	0•18	0•06		0•28	0•18	0•20	0•19	0•32	0•31	0•32	0•35	0•34	0•44	0.54
4	Pays-Bas (ouvriers)	0•22	0•40	0•28		0•17	0•17	0•33	0•45	0•25	0•26	0.55	0•46	0.53	0.55
5	Hamburg (ouvriers)	0•27	0•25	0•18	0•17		0•07	0•29	0•40	0•26	0•23	0•49	0•47	0•45	0.50
6	Deutschland (- Hamburg) (Ouvr.)	0•37	0•37	0•20	0•17	0•07		0•31	0•41	0•21	0•19	0.56	0•47	0•46	0•43
7	Belgique (Employés, fonction.)	0•28	0•26	0•19	0•33	0•29	0•31		0•23	0•13	0•21	0•20	0•16	0•31	0•37
8	Luxembourg (Employés, fonction.)	0•47	0•39	0•32	0•45	0•40	0•41	0•23		0•32	0•25	0•40	0•34	0•46	0.54
9	Belgique (Ouvriers)	0•49	0•48	0•31	0•25	0•26	0•21	0•13	0•32		0•12	0•42	0•27	0•31	0•18
10	Luxembourg (Ouvriers)	0•49	0•47	0•32	0•26	0•23	0•19	0•21	0•25	0•12		0•37	0•26	0•22	0•28
11	Paris (Employés, fonction.)	0•42	0•34	0•35	0.55	0•49	0.56	0•20	0•40	0•42	0•37		0•09	0•16	0•47
12	France (- Paris) (Empl. Fonct.)	0•46	0•41	0•34	0•46	0•47	0•47	0•16	0•34	0•27	0•26	0•09		0•14	0•28
13	Paris (Ouvriers)	0.64	0.52	0•44	0.53	0•45	0•46	0•31	0•46	0•31	0•22	0•16	0•14		0•24
14	Nord (Ouvriers)	0.84	0.73	0.54	0.55	0.50	0•43	0•37	0.54	0•18	0•28	0•47	0•28	0•24	
15	Ouest (Ouvriers)	0.83	0.74	0.56	0.60	0.57	0•50	0•39	0.53	0•26	0•29	0•38	0•19	0•20	0•11
16	Massif-Central (Ouvriers)	0.86	0.75	0.57	0.59	0.59	0•50	0•44	0.54	0•32	0•29	0•40	0•23	0•17	0•17
17	Sud-Ouest (Ouvriers)	0.96	0.86	0.67	0.71	0.69	0.60	0.54	0.63	0•43	0•39	0•49	0•30	0•27	0•31
18	Bas.Paris. Est,Sud-Est,Midi	0.68	0.61	0•45	0•44	0•45	0•39	0•30	0•43	0•20	0•20	0•29	0•12	0•11	0•11
19	Italie du Nord (Empl. fonction)	0.62	0.62	0.53	0.52	0•46	0•49	0•34	0.51	0•40	0•30	0•29	0•23	0•26	0•48
20	Italie du Sud (Empl. fonction)	0.86	0.84	0.69	0.65	0.59	0.60	0.54	0.68	0.51	0•41	0.58	0•44	0•45	0.63
21	Italie du Centre (Empl. fonct.)	0.82	0.77	0.65	0.67	0.58	0.61	0•47	0.60	0.51	0•37	0•45	0•37	0•39	0.65
22	Piemonte, Aoste, Liguria (Ouv.)	0.93	0.89	0.72	0.62	0.56	0.56	0.61	0.68	0•47	0•32	0.61	0•47	0•35	0•47
23	Lombardia,Tre Venezia,Emil. Mar	0.90	0.85	0.67	0.59	0.53	0•48	0.53	0.66	0•38	0•32	0.57	0•42	0•32	0•38
24	Toscana,Umbria,Alto Lazio (Ouv)	1.09	1.04	0.84	0.78	0.70	0.66	0.69	0.81	0.54	0•48	0.74	0.56	0•46	0.51
25	Lazio Meridionale,Campania (Ouv)	1.53	1.45	1.22	1.12	1.02	0.98	1.13	1.24	0.91	0.81	1.24	1.05	0.91	1.00
26	Abruzzi e Molizi (Ouvriers)	1.44	1.33	1.10	1.06	0.93	0.87	1.00	1.11	0.78	0.69	1.09	0.88	0.75	0.74
27	Puglia,Calab.Basil.Sicil.Sarde.	1.36	1.31	1.08	0.97	0.92	0.85	0.95	1.07	0.75	0.67	1.08	0.85	0.77	0.83
28	Roma e Provincia.	1.16	1.07	0.88	0.84	0.73	0.71	0.74	0.85	0.61	0•50	0.81	0.68	0.56	0.72

1 - Le terme général de cette matrice symétrique est le carré de la distance entre deux pays i et i' : $d^2(i,i)$ telle qu'elle est définie page 2.

Annexe IV - (suite)

	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Pays-Bas (Employés,fonction.)	0.83	0.86	0.96	0.68	0.62	0.86	0.82	0.93	0.90	1.09	1.53	1.44	1.36	1.16
Hamburg " "	0.74	0.75	0.86	0.61	0.62	0.84	0.77	0.89	0.85	1.04	1.45	1.33	1.31	1.07
Deutschland (-Hamburg) "	0.56	0.57	0.67	0*45	0.53	0.69	0.65	0.72	0.67	0.84	1.22	1.10	1.08	0.88
Pays-Bas (Ouvriers)	0.60	0.59	0.71	0*44	0.52	0.65	0.67	0.62	0.59	0.78	1.12	1.06	0.97	0.84
Hamburg (Ouvriers)	0.57	0.59	0.69	0*45	0*46	0.59	0.58	0.56	0.53	0.70	1.02	0.93	0.92	0.73
Deutschland (-Hamburg) (Ouv.)	0*50	0*50	0.60	0*39	0*49	0.60	0.61	0.56	0*48	0.66	0.98	0.87	0.85	0.71
Belgique (Employés,fonction.)	0*39	0*44	0.54	0*30	0*34	0.54	0*47	0.61	0.53	0.69	1.13	1.00	0.95	0.74
Luxembourg (Employés,fonct.)	0.53	0.54	0.63	0*43	0.51	0.68	0.60	0.68	0.66	0.81	1.24	1.11	1.07	0.85
Belgique (Ouvriers)	0*26	0*32	0*43	0*20	0*40	0.51	0.51	0*47	0*38	0.54	0.91	0.78	0.75	0.61
Luxembourg (Ouvriers)	0*29	0*29	0*39	0*20	0*30	0*41	0*37	0*32	0*32	0*48	0.81	0.69	0.67	0*50
Paris (Employés,fonction.)	0*38	0*40	0*49	0*29	0*29	0.58	0*45	0.61	0.57	0.74	1.24	1.09	1.08	0.81
France (-Paris) (Empl.fonct.)	0*19	0*23	0*30	0*12	0*23	0*44	0*37	0*47	0*42	0.56	1.05	0.88	0.85	0.68
Paris (Ouvriers)	0*20	0*17	0*27	0*11	0*26	0*45	0*39	0*35	0*32	0*46	0.91	0.75	0.77	0.56
Nord (Ouvriers)	0*11	0*17	0*31	0*11	0*48	0.63	0.65	0*47	0*38	0.51	1.00	0.74	0.83	0.72
Ouest (Ouvriers)	[redacted]	0*13	0*23	0*08	0*40	0.51	0.53	0*45	0*36	0*47	0.94	0.70	0.75	0.67
Massif-Central (Ouvriers)	0*13	[redacted]	0*18	0*08	0*40	0*50	0.50	0*35	0*31	0*39	0.85	0.64	0.67	0.59
Sud-Ouest (Ouvriers)	0*23	0*18	[redacted]	0*14	0*37	0*38	0*43	0*31	0*25	0*31	0.60	0*46	0*48	0*45
Bas.Paris,Est,Sud-Est,Midi (O)	0*08	0*08	0*14	[redacted]	0*28	0*39	0*39	0*31	0*24	0*36	0.77	0.61	0.60	0.50
Italie du Nord (Empl.fonct.)	0*40	0*40	0*37	0*28	[redacted]	0*14	0*09	0*20	0*16	0*25	0.58	0.51	0*49	0*29
Italie du Sud (Empl.fonction.)	0.51	0*50	0*38	0*39	0*14	[redacted]	0*08	0*16	0*15	0*15	0*24	0*25	0*18	0*13
Italie du Centre (Empl.fonct.)	0.53	0.50	0*43	0*39	0*09	0*08	[redacted]	0*17	0*20	0*24	0*39	0*42	0*36	0*13
Piemonte,Aosta,Liguria (Ouv.)	0*45	0*35	0*31	0*31	0*20	0*16	0*17	[redacted]	0*09	0*12	0*29	0*23	0*27	0*13
Lombardia,Tre Venezia,Emil.Mar	0*36	0*31	0*25	0*24	0*16	0*15	0*20	0*09	[redacted]	0*08	0*31	0*23	0*24	0*16
Toscana,Umbria,Alto Lazio (Ouv)	0*47	0*39	0*31	0*36	0*25	0*15	0*24	0*12	0*08	[redacted]	0*24	0*17	0*20	0*15
Lazio Meridionale,Campania (Ou)	0.94	0.85	0.60	0.77	0.58	0*24	0*39	0*29	0*31	0*24	[redacted]	0*12	0*06	0*14
Abruzzi e Molizi (Ouvriers)	0.70	0.64	0*46	0.61	0.51	0*25	0*42	0*23	0*23	0*17	0*12	[redacted]	0*13	0*22
Puglia,Calab.Basil.Sicil.Sarde	0.75	0.67	0*48	0.60	0*49	0*18	0*36	0*27	0*24	0*20	0*06	0*13	[redacted]	0*18
Roma e Provincia (Ouvriers)	0.67	0.59	0*45	0.50	0*29	0*13	0*13	0*13	0*16	0*15	0*14	0*22	0*18	[redacted]


ANNEXE 5
COMMUNAUTE ECONOMIQUE EUROPEENNE
BUDGETS FAMILIAUX DES SALARIES
Analyse des correspondances :
- 55 Unités socio-géographiques
- 67 Postes de consommation
Facteurs 1 et 3
F3 (11,1%)
F1 (41,3%)
ITALIE : Employés, Fonctionnaires
ITALIE : Ouvriers
FRANCE : Employés fonctionnaires
FRANCE : Ouvriers
BELGIQUE : Ouvriers
LUXEMBOURG
ALLEMAGNE : Employés Fonctionnaires
ALLEMAGNE : Ouvriers
PAYS-BAS : Employés Fonctionnaires
PAYS-BAS : Ouvriers
Abruzzi
Roma
Piemonte
Emilia
Sicilia
Puglia
Campania
Toscana
Lombardia
Tre Venezia
Paris
Sud-Est
Sud-Ouest
Massif Central
Est
Bassin Parisien
Ouest
Nord
Baden
Hessen
NordRhein
Bayern
Schleswig
Rheinland
Saarland
Hamburg
Huile
Riz, pâtes
Hôtel, Logement vacances
Assurances mobilières
P et T. Déménagements
Enseignement
Réparations logement, chauffage
Transp. publics
Spectacle vacances
Electricité
Loyer et Charges
Gaz
Fruits
Tissus façon
Chaussures
Légumes
Agrumes
Bananes
Pain
Sucre
Méditerranée
Télévision
Volailles
Poissons alcoolisés
Livres, journaux périodiques
Ameublement
Utilis. de véhicules
Véhicules
Réparations Chaussures
Coiffure, beauté
Articles ménagers
Confiserie, confiture
Textiles d'ameublement
Charcuterie
Poissons sans alcool
Farines Céréales
Pommes de terre
Combustibles liquides et solides
Beurre
Margarine
Porc

EXEMPLE 3

III - STRUCTURE DE LA POPULATION D'UN ECHANTILLON DE COMMUNES SELON LE TRAVAIL FEMININ ET LE NOMBRE D'ENFANTS

III-1. Présentation des variables analysées.

L'analyse porte sur l'ensemble des communes ou arrondissements retenus pour le sondage de l'enquête CNAF 1971, soit 104 unités géographiques dites ici "communes-échantillon".

L'échantillon de cette enquête a été construit de façon à obtenir une représentation identique des familles quel que soit le nombre des enfants et le fait que la mère travaille ou non. Pour le tirage et le redressement de cet échantillon, on a donc obtenu des Caisses d'Allocations Familiales les statistiques par commune des familles allocataires selon les trois variables suivantes croisées entre elles :

1 - perception ou non du salaire unique
2 - nombre d'enfants allocataires de 2 à 5
3 - âge de l'aîné des enfants allocataires en 4 classes (0 - 5 ans, 6 - 9 ans, 10 - 14 ans, 15 ans et plus).

Le croisement de ces critères conduit à une partition en 32 classes des familles.

Nous analysons ici un tableau de contingence de 104 lignes et 36 colonnes, donnant la répartition des 588 609 familles ayant de 2 à 5 enfants allocataires (1), selon leur commune de résidence i d'une part (i = 1,...., 104) et leur type de famille j d'autre part.

---

(1) On compte, pour l'ensemble des communes, une population totale de 2.980.728 familles allocataires, du régime général salariés, au 31.12.1970. L'analyse présentée ici se rapporte donc à une fraction importante de la population totale. Comme le choix des villes d'enquête a été fait de façon à représenter tous les profils d'activités et de professions, on peut penser que la structure qui se dégage de l'analyse des tableaux ci-après est très voisine de celle que l'on obtiendrait à partir de l'information exhaustive.

Ce tableau est beaucoup trop volumineux pour figurer dans un rapport. Seule sa plus petite marge est présentée plus loin, qui fournit, en même temps que des ordres de grandeurs, la liste des 32 variables retenues.

L'essentiel de l'information du tableau initial est contenu dans le graphique 1 : le premier plan d'inertie exprime 64% de sa dispersion totale (40,16% pour le premier axe, 23,67% pour le second).

III-2. Interprétation des résultats.

La première chose qui frappe dans ce graphique est l'inclinaison par rapport aux axes de la répartition des variables : le nombre d'enfants augmente dans une direction approximativement parallèle à la première bissectrice (1) et la coupure entre salaire unique et non-salaire unique se fait en direction de la seconde bissectrice. Les deux principaux caractères qui composent les variables de départ : travail féminin et nombre d'enfants ne sont pas indépendants des zones géographiques.

Le premier axe est ainsi fonction des deux à la fois ; il oppose les familles ayant peu d'enfants et dans lesquelles la mère travaille, à gauche sur F1, aux familles ayant beaucoup d'enfants et où la mère reste au foyer, à l'extrême droite.

Et cependant l'analyse des correspondances fait bien ressortir distincts les deux caractères travail féminin et nombre d'enfants. De telle sorte qu'on peut repérer l'existence dans la réalité de tous les cas possibles : il y a des mères de famille nombreuse qui travaillent, des mères qui ne travaillent pas et ont peu d'enfants.

Ces cas sont plus rares, il s'opposent sur le second axe.

La configuration du premier plan est alors très claire. Elle exprime toutes les combinaisons des deux traits analysés selon leur importance dans la réalité, lorsqu'on passe d'une ville à l'autre sur ce plan on connait immédiatement leur différence.

---

(1) On notera les variations selon l'âge de l'aîné : plus l'aîné est jeune (familles en probabilité non terminées), plus on se rapproche des familles du rang supérieur.

A Villejuif, Saint-Ouen, Châtillon il y a peu de familles nombreuses et les femmes travaillent. A Châtillon par exemple on compte 53% de familles de 2 enfants (contre 44% pour l'ensemble), 26% ne touchent pas le salaire unique (contre 17%).

En ligne oblique vers le haut, le nombre d'enfants augmente : Ivry, Sarcelles, Saint-Denis, Gennevilliers, La Courneuve... et les femmes travaillent toujours.

En ligne oblique vers le bas le nombre d'enfants n'augmente pas, Saint-Maur des Fossés, Paris 16ème, Saint-Cloud, Le Vésinet, Nice et Béziers, mais les familles où la mère travaille sont de plus en plus rares.

Il n'y a pas de raison a priori pour que les proximités entre villes du point de vue de leurs structures familiales coïncident avec des proximités géographiques. Il se trouve cependant qu'il y a un lien assez fort entre les deux phénomènes : des zones régionales se dessinent sur le graphique 1.

Le premier axe oppose la zone la plus industrielle de la région parisienne (taux d'activité féminine élevé, relativement peu de familles nombreuses) à la région du Nord (1) qui a les caractéristiques opposées. La ville de Brest se trouve au milieu des communes du Nord avec les mêmes caractéristiques que celle -ci... Le second axe semble opposer la région de l'Ouest à la vallée du Rhône jusqu'au midi méditerranéen.

L'effet géographique est beaucoup plus clair si l'on applique les résultats de l'analyse des correspondances sur la carte de France elle-même.

Deux axes recourbés partagent l'ensemble de la France :

- le premier, qui correspond à peu près à la première bissectrice du graphique 1 va des Pyrénées Orientales au Pas-de-Calais, en décrivant un arrondi (les deux axes se rencontrent vers l'Aube). Il sépare grossièrement l'Ouest, où le taux d'activité féminine est le

---

(1) Des petites communes du département du Nord ont été regroupées. Elles figurent sur le graphique 1 avec la désignation "59-groupement de communes 1, 2 ou 3".

plus élevé ($F1 < 0$, $F2 > 0$) à l'Est où il est le plus faible.

- le second va de Bordeaux à Metz. Il correspond à peu près à la seconde bissectrice du graphique 1. La partie supérieure de cet axe est la zone où la fréquence de familles nombreuses est la plus élevée : Nord, Nord-Ouest (F1 et F2 sont positifs).

La région parisienne fait exception. A cette partition géographique, se superpose une proximité entre les métropoles régionales caractérisée par un taux d'activité féminine fort et une natalité faible : région lyonnaise, région de Bordeaux, de Rouen, de Toulouse, mais ce trait est beaucoup moins marqué que le caractère géographique (ces métropoles sont assez proches de l'origine sur le graphique 1).

F2
23.67%

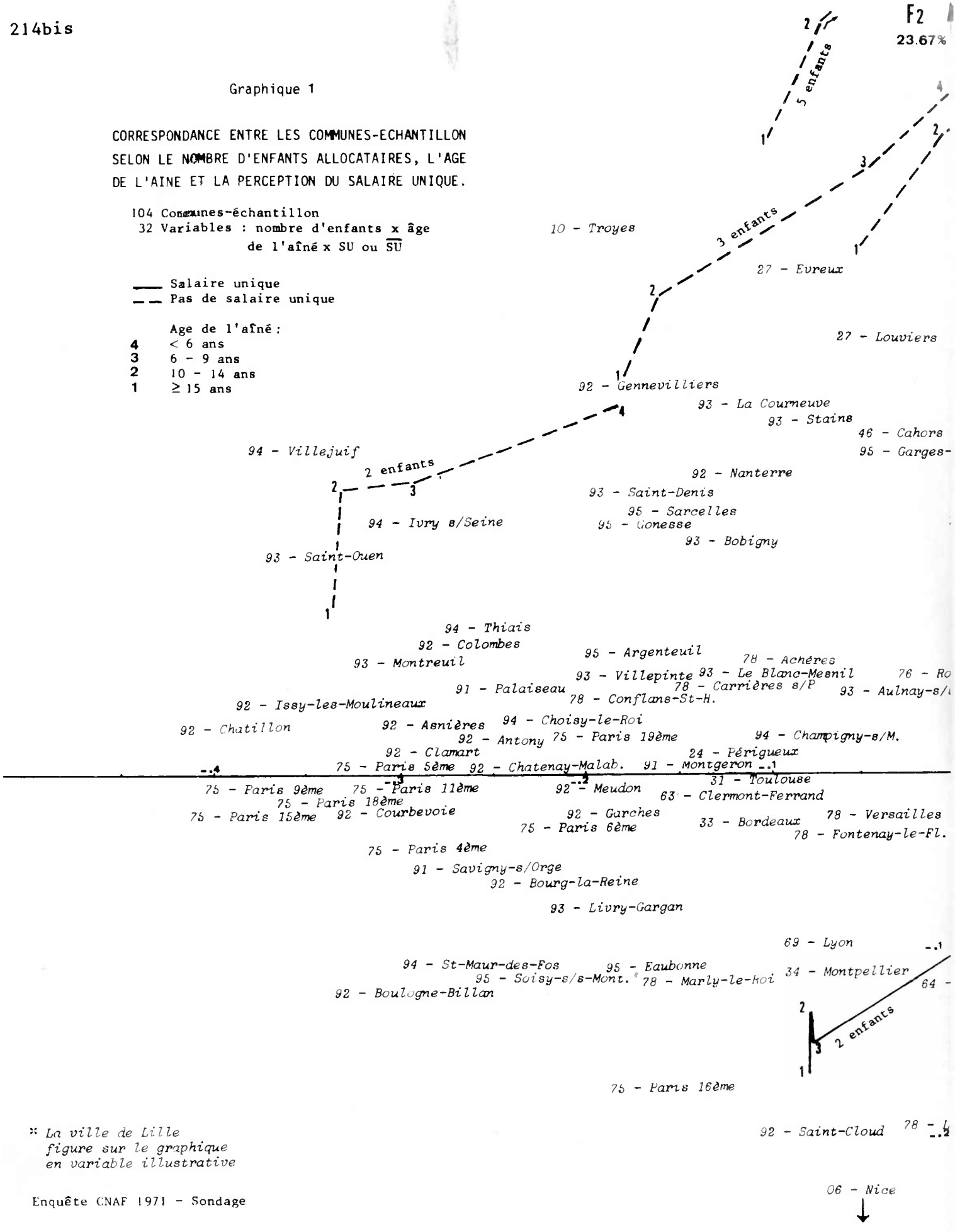
Graphique 1

CORRESPONDANCE ENTRE LES COMMUNES-ECHANTILLON SELON LE NOMBRE D'ENFANTS ALLOCATAIRES, L'AGE DE L'AINE ET LA PERCEPTION DU SALAIRE UNIQUE.

104 Communes-échantillon
32 Variables : nombre d'enfants x âge de l'aîné x SU ou $\overline{SU}$

—— Salaire unique
– – Pas de salaire unique

Age de l'aîné :
4 < 6 ans
3 6 - 9 ans
2 10 - 14 ans
1 ≥ 15 ans



* *La ville de Lille figure sur le graphique en variable illustrative*

Enquête CNAF 1971 - Sondage

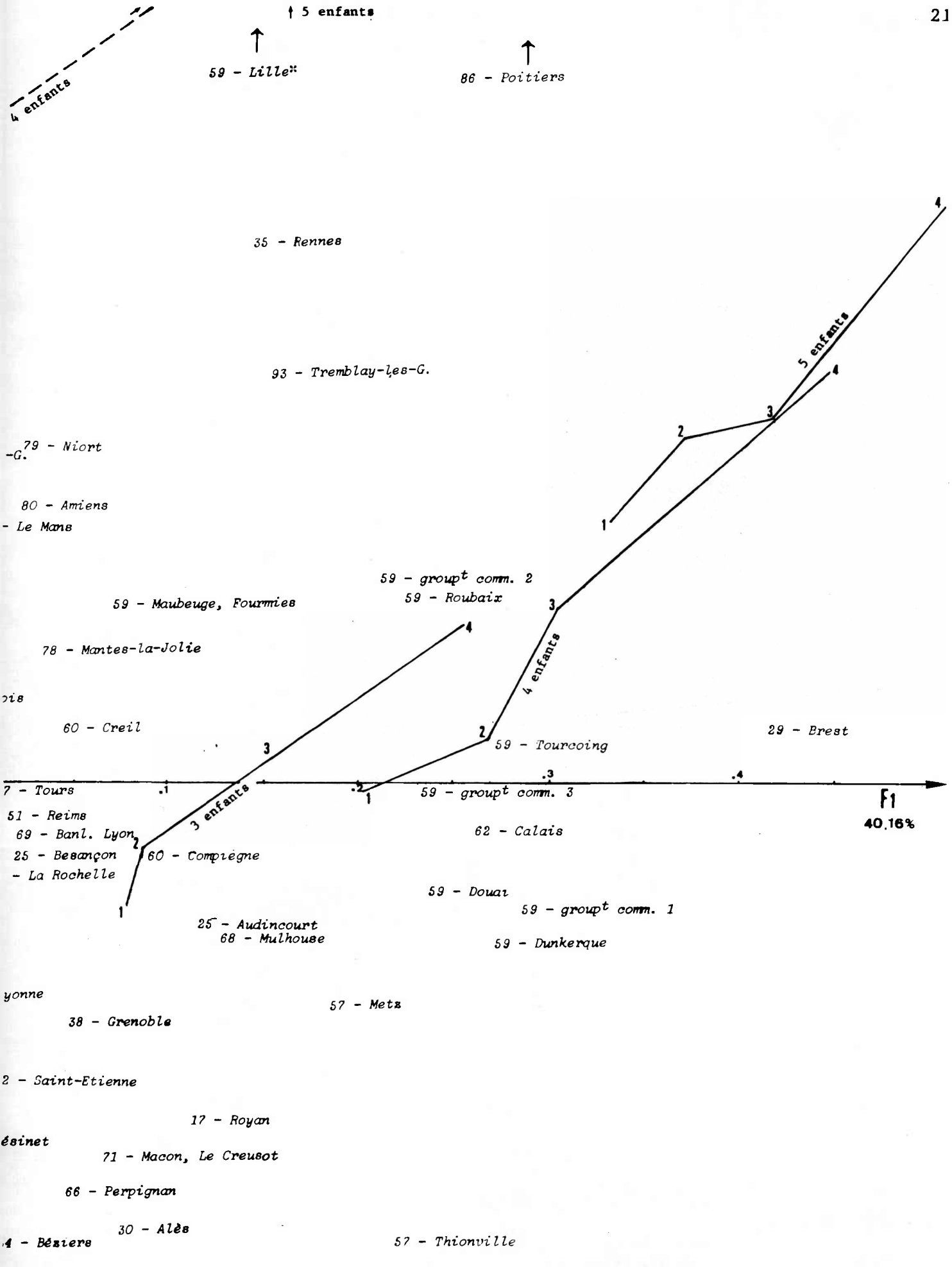

5 enfants
4 enfants
59 - Lille
86 - Poitiers
35 - Rennes
93 - Tremblay-les-G.
79 - Niort
80 - Amiens
- Le Mans
59 - groupt comm. 2
59 - Roubaix
59 - Maubeuge, Fourmies
78 - Mantes-la-Jolie
60 - Creil
59 - Tourcoing
29 - Brest
7 - Tours
51 - Reims
69 - Banl. Lyon
25 - Besançon
- La Rochelle
3 enfants
60 - Compiègne
59 - groupt comm. 3
62 - Calais
59 - Douai
59 - groupt comm. 1
25 - Audincourt
68 - Mulhouse
59 - Dunkerque
yonne
57 - Metz
38 - Grenoble
2 - Saint-Etienne
17 - Royan
ésinet
71 - Macon, Le Creusot
66 - Perpignan
30 - Alès
4 - Béziers
57 - Thionville
F1
40.16%

TABLEAU 1

POPULATION TOTALE DES VILLES ENQUETEES

nombre et pourcentage des familles

Familles percevant le salaire unique

Age de l'aîné	Nombre d'enfants				
	1	2	3	4	5
0 à 5 ans	80.192 *	77.125	17.808	3.106	456
	1.069	1.028	237	41	6
6 à 9 ans	11.600 *	50.714	34.146	12.683	4.776
	155	676	455	169	64
10 à 14 ans	30.260 *	57.713	42.809	22.479	11.382
	404	770	571	300	152
15 ans et +	39.272 *	55.377	38.499	20.994	11.432
	524	738	513	280	153

Ensemble 622.845 familles (8.305°/$_{000}$)

Familles ne percevant pas le salaire unique

Age de l'aîné	Nombre d'enfants			
	2	3	4	5
0 à 5 ans	22.189	2.068	168	14
	296	28	2	1
6 à 9 ans	21.840	5.381	932	257
	291	72	12	3
10 à 14 ans	25.707	8.699	2.642	1.177
	343	116	35	16
15 ans et +	22.751	8.779	3.004	1.480
	303	117	40	10

Ensemble 127.088 familles (1.695°/$_{000}$)

Ensemble : 749.933 familles dont 588.609 ayant de 2 à 5 enfants

1ère ligne : nombre de familles allocataires
2ème ligne : répartition pour 10.000 familles

Les cases munies d'un astérisque ont été supprimées pour l'analyse, par raison de symétrie entre les familles percevant le salaire unique et celles ne le percevant pas.

ANNEXE

NOTE SUR L'ANALYSE INTERACTIVE DES DONNEES STATISTIQUES

Exemples d'utilisation du langage "A P L" [1]

Les méthodes d'analyse de données, ou encore de statistique descriptive multidimensionnelle, qui se sont développées sous l'impulsion du Professeur BENZECRI, doivent leur existence à l'apparition du calcul automatique, c'est-à-dire à une innovation technologique. Les applications licites de ces méthodes aux sciences humaines permettent de réaliser d'incontestables progrès dans la connaissance des domaines les plus variés, mais la diffusion rapide de ces techniques n'est pas sans poser de problèmes d'information, de communication. C'est pourquoi toute innovation à caractère technique ayant pour effet de rendre plus accessible l'ordinateur matériellement et psychologiquement, doit vraisemblablement avoir des répercussions sur le mode d'utilisation de ces méthodes.

Le langage de programmation conversationnel "A P L", mis au point au cours des années soixante par le Dr Kenneth IVERSON constitue une innovation de ce type. A la fois plus puissant et plus simple que la plupart des langages couramment utilisés actuellement, ce langage qui a déjà des dizaines de milliers d'utilisateurs aux Canada et aux Etats-Unis, permet un dialogue "statisticien-ordinateur" remarquable de clarté et de concision.

Cette note a pour objet la présentation succincte des diverses caractéristiques de ce langage ainsi que des avantages qui résultent de son utilisation à propos de problèmes d'analyse de données. La présentation du langage s'adresse aux lecteurs non initiés, qui doivent en retirer une impression générale leur permettant de situer les traits les plus remarquables d'"A P L" par rapport aux autres langages. Cette présentation ne peut évidemment se substituer à un cours ou à un stage de formation.

Les remarques et les exemples qui suivent sont également destinés au plus grand nombre de lecteurs statisticiens, mais l'annexe intéressera plus spécialement le lecteur déjà un peu familier du langage "A P L", ou alors le lecteur non initié qui aurait consulté entre temps les ouvrages cités en bibliographie.

Avant la présentation proprement dite, donnons quelques caractéristiques pratiques des modalités d'utilisation des systèmes conversationnels du type de ceux qui nous intéressent ici :

L'utilisateur dispose d'un terminal qui peut être branché sur n'importe quel poste téléphonique. Il entre en contact avec l'ordinateur à la suite d'un simple appel. Il frappe alors sur le clavier du terminal les instructions qu'il désire voir exécuter, et donne la parole à l'ordinateur en frappant, par "retour-chariot". Celui-ci donne une réponse, signale une erreur, ou indique qu'il attend une autre instruction, selon les cas. La réponse de la machine se distingue, sur le papier enregistreur, des instructions de l'utilisateur, par un décalage, ou, pour certaines installations, par une couleur différente. S'il le désire, et s'il a le téléphone, l'utilisateur peut emporter le terminal chez lui pour terminer le travail en cours. Tout ceci est fort simple, et dans une certaine mesure, démystificateur. Beaucoup d'intermédiaires sont supprimés.

---

1 - Rédigé avec la collaboration de G. LACOURLY

Ce mode de connection avec la machine sera rentable si le langage laisse certaines libertés à l'utilisateur : celui-ci doit avoir le droit de se tromper de temps en temps, de ne pas avoir tout prévu à l'avance, d'écrire les choses à peu près comme il les pense.

Les qualité du langage "A P L" vont permettre un tel contact direct avec la machine. Le mode conversationnel, par le dialogue et l'autorégulation qu'il implique, possède des avantages psychologiques indéniables : la sanction immédiate de la machine permet un apprentissage extrêmement rapide, qu'il s'agisse au début d'apprendre le langage lui-même, ou plus tard de faire face à des situations variées (mise au point d'algorithmes, analyse de données statistiques). Ces qualités pédagogiques ne sont pas à négliger puisque notre domaine d'application, les sciences humaines et économiques, suppose des contacts permanents entre personnes de formations très différentes. L'utilisation d'un ordinateur comme prolongement d'un calculateur de bureau, sans la mise à contribution de services et de spécialistes parfois mystérieux, est propre à dissiper bien des malentendus.

En bref, toute innovation susceptible de faire tomber des cloisonnements dans un domaine d'activité interdisciplinaire mérite un accueil, ou au moins un regard favorable.

Examinons maintenant quelles sont les opérations pratiques auxquelles se livre le statisticien désireux d'analyser le contenu de quelques tableaux rectangulaires de données. Pour les tableaux de petites dimensions, il se contente de les lire directement, en calculant éventuellement quelques paramètres sur une machine de bureau (moyennes, totaux marginaux, etc...) ; pour les tableaux de dimensions moyennes, il hésite à mettre en branle le service de calcul, ou à programmer et perforer les instructions nécessaires lui-même. Ceci suppose en effet des démarches et des délais qui peuvent être gênants pour la poursuite de l'étude en cours. Les tableaux importants seront peut-être passés de façon plus systématique à "la moulinette", mais les premiers résultats révèleront que tel codage est défectueux, que deux groupes d'individus gagneraient à faire l'objet d'analyses séparées, qu'une erreur s'était glissée dans le fichier. Ceci suppose des allers et retours, des contacts, en un mot, une mobilisation du chercheur sur un travail de programmation ou de gestion qui peut le détourner pendant un certain temps des problèmes spécifiques de sa discipline.

A ces va-et-vient parfois accidentels s'ajoutent les tatonnements inhérents aux méthodes de statistique descriptive : on désire illustrer une représentation par l'adjonction de variables supplémentaires, effectuer une nouvelle transformation préalable des données, vérifier la validité des résultats par une simulation séquentielle qui permet de faire émerger progressivement les réseaux d'associations significatives. Toutes ces opérations supposent les résultats connus à chaque étape ; elles sont grandement facilitées par une connection directe entre le statisticien et l'ordinateur.

I - PRESENTATION SOMMAIRE DU LANGAGE A P L

I-1. Caractéristiques générales.

Mis au point par K.E. IVERSON et A.D. FALKOFF, le langage de programmation A P L est principalement destiné à l'exécution de calculs scientifiques en mode conversationnel.

Une des caractéristiques fondamentales de ce langage est l'existence d'un ensemble complet d'opérateurs sur tableau. L'allocation de ces tableaux est dynamique (il n'est pas besoin de spécifier à l'avance les dimensions maxima de ces tableaux). Le foisonnement des opérateurs a conduit à adopter une règle de priorité qui peut paraître surprenante au prime abord : un opérateur s'appliquera à tout ce qui est à sa droite. Ainsi, 1 x 2 + 3 signifiera 1 x (2 + 3). En fait, cette règle apparait bien vite comme étant très féconde, et peu astreignante.

Le plus simple, pour établir un premier contact avec le langage A P L est de s'installer devant le clavier de la machine à écrire du terminal, et de tenter de donner quelques instructions :

Commençons par l'utilisation en calculateur de bureau ; frappons sur le clavier.

```
      3.1 x 4
12.4
```

l'ordinateur nous a répondu, avec un léger décalage sur la gauche, afin de nous permettre d'identifier les lignes de questions et les lignes de réponses.

Essayons maintenant :

```
A ← 3 5 7
```

Pas de réponse. L'ordinateur s'est contenté d'affecter à une zone mémoire qui s'appellera dorénavant A le vecteur à trois composantes 3 5 7. On notera que les "blancs" jouent un rôle important, et qu'il n'est pas nécessaire d'utiliser d'indice pour définir le vecteur A. Si nous voulons vérifier le contenu de la mémoire A, il suffit de frapper la lettre A :

```
      A
3 5 7
```

On remarque qu'il n'a pas été besoin de spécifier le type de A (entier, réel), ni de spécifier un modèle d'impression.

Si nous frappons maintenant :

```
      A+2
5 7 9
```

La réponse nous indique que le nombre 2 a été ajouté à chacune des composantes du vecteur A. Ceci est une règle générale en A P L : les opérateurs dyadiques tels que l'addition se généralisent aux tableaux de la façon suivante : si les tableaux ont les mêmes dimensions, l'opérateur agit sur les couples d'éléments terme à terme. Si l'on a un tableau et un scalaire, comme dans notre exemple, l'opérateur agit pour chaque couple formé d'un élément du tableau et du scalaire. Enfin, si l'on a affaire à deux tableaux de dimensions différentes, une erreur est signalée, sans exécution.

Continuons nos essais :

```
      ρA
3
```

Nous avons demandé, à l'aide du symbole ρ , la dimension du tableau A. Cette dimension peut être un vecteur, si A est une matrice, ou un tableau à n dimensions.

```
      +/A
15
```

L'opérateur / placé derrière un signe d'opération quelconque correspond à l'insertion de ce signe entre tous les éléments du tableau.

On a donc effectué : 3 + 5 + 7 = 15

```
      (+/A) ÷ ρA
5
```

Le signe ÷ étant celui de la division, nous avons divisé la somme des éléments de A par le nombre de ses éléments. Ces instructions simples nous permettent de calculer la moyenne arithmétique des éléments d'un tableau sans même connaître le nombre d'éléments de ce tableau.

Construisons une fonction qui calcule cette moyenne.

```
∇ MOYENNE A
[1] (+/A)÷ ρA
[2] ∇
```

Lorsque nous avons frappé le triangle à l'envers, le système a été averti de notre intention de construire une fonction. Il a numéroté lui-même les instructions en marge, et s'arrête lorsque apparaît un second triangle.

Essayons le programme MOYENNE.

```
MOYENNE 2 3 4 5 6
4
```

Il est certes plus long d'écrire le mot MOYENNE que d'écrire les instructions de calcul nécessaires. Indépendamment de l'intérêt pédagogique de cet exemple, l'utilisateur statisticien aura quand même

avantage à construire de cette façon un sous-langage adapté à sa sphère d'intérêt.

Parmi les opérateurs scalaires dyadiques, il y a, bien sûr, tous ceux existant dans les autres langages (addition, soustraction, division, multiplication, exponentiation, logarithme dans une base donnée, résidu de la division de B par C, opérations logiques : et, ou, identité, etc...), l'opérateur "plafond" qui donne le plus grand A⌈B de deux éléments A et B, l'opérateur "plancher", qui donne le plus petit A⌊B de deux éléments A et B.

Si A et B sont des matrices de mêmes dimensions, avec la règle énoncée plus haut, A x B désigne le produit contracté des matrices A et B. Le produit matriciel ordinaire existe, lorsque les dimensions des matrices mises en jeu s'y prètent. On le note : A + . x B.

Ce produit peut être généralisé, en remplaçant les signes + et x par n'importe quels signes, pouvu que ceux-ci soient relatifs à des opérateurs dyadiques. Ainsi, le produit A⌈.xB nous donne une matrice dont le terme général est le plus grand élément des produits terme à terme de chaque ligne de A par chaque colonne de B.

Indépendamment des produits contractés, du produit matriciel généralisé, il existe un produit "extérieur", pour deux matrices de dimensions quelconques, qui effectue une opération dyadique quelconque entre tout élément de A et tout élément de B. Si l'opération est x, on obtient des produits tensoriels. On le note dans ce cas ∘.x . On peut construire de la même façon ∘.+, ∘.⌈, etc... Par exemple (1 2)∘.+(3 5) donne la matrice :

```
4 6
5 7
```

D'autres opérations sur tableaux sont également cablées : ainsi, l'opérateur ⍋ appliqué à un vecteur donne les rangs des éléments de ce vecteur, lorsque ceux-ci sont classés par ordre croissant.

L'inverse de la matrice A se note ⌹ A, le tirage de 100 nombres au hasard entre 1 et 10000 se note 100?10000.

Ces quelques informations suffisent à donner une idée, encore vague certes, des possibilités du langage A P L. Ses caractéristiques sont particulièrement favorables aux calculs statistiques usuels, aux calculs portant sur les rangs, à toutes les procédures utilisées en algèbre linéaire. De toute façon, le système de construction des fonctions est tellement souple que le langage s'adapte facilement ; néanmoins, les opérateurs et fonctions "cablés" sont extrêmement performants.

I-2. Modalités d'implantation du langage A P L

A P L est réalisé dans un environnement de temps partagé, c'est-à-dire que l'utilisateur est en liaison directe avec l'ordinateur, et bien que n'étant pas le seul dans ce cas, a l'impression de disposer de ce dernier continûment et en totalité. A toute question et demande d'exécution de programme, l'ordinateur répond instantanément, sauf évidemment si le calcul demandé mobilise l'unité centrale pendant un temps important.

Les calculs, les définitions de fonctions, les enregistrements de données se font dans une zone appelée "Espace actif de travail" (la taille de cette zone est variable selon les installations) elle est souvent de l'ordre de 32.000 octets, soit 4000 mots de 64 bits . L'utilisateur dispose de plusieurs espaces de travail stockés sur des disques avec lesquels il peut communiquer. Il peut à tout moment copier son espace de travail actuel sur disque, afin de préserver ses résultats, charger un autre espace de travail, copier une fonction ou un fichier situé sur un autre espace de travail, etc...

Bien que les transactions entre espaces de travail soient très simples, il n'en existe pas moins une limite en taille. En effet, on ne peut créer ou copier dans l'espace actif de travail un tableau de taille supérieure à celle de cet espace actif. C'est alors qu'il est nécessaire de disposer d'un système de fichiers. Le système de fichiers doit être simple et indépendant de la machine et de ses ressources. C'est pourquoi la logique du système doit être aussi simple que son utilisation. Pour l'utilisateur, un fichier est un ensemble de composants, chacun d'eux pouvant être une entité A P L, c'est-à-dire un scalaire, un tableau à n dimensions, une chaîne de caractères.

Pour avoir accès au composant 3 du fichier appelé par exemple LOISIR, il suffit de frapper :

```
LIRE LOISIR 3
```

Pour modifier ou ajouter des éléments il existe des instructions aussi simples. Pour utiliser des fichiers volumineux, il suffit donc d'utiliser des fonctions de gestion de fichier, comme d'autres fonctions A P L.

II - EXEMPLE PRATIQUE D'ANALYSE DE TABLEAU STATISTIQUE

Nous donnons ci-dessous un exemple pratique d'introduction de données, d'appel de sous programme, enfin d'édition de résultats sur la machine à écrire du terminal. Les programmes utilisés figurent dans leur intégralité en annexe. Les données analysées ayant fait l'objet d'un article[1], nous nous limiterons à des considérations méthodologiques, sans insister sur l'interprétation économique des représentations obtenues.

Une petite fonction, construite pour l'occasion, nous permet de charger le tableau de données à partir de la machine à écrire, sans grand risque d'erreur (cf. page). Le tableau, de dimensions 15 x 31, nous donne pour les 31 régions d'Europe précédemment étudiées les valeurs de 15 consommations par ménage ouvrier.

---

1 - Cf. Chapitre IV, § 2.

Il nous faut introduire ensuite des identificateurs de chacune de ces consommations et de chacune de ces régions : nous chargeons alors un tableau de dimensions 5 x 46 caractérisant par 5 lettres les (31 + 15) individus et variables. Nous choisissons des abréviations pour les régions, faisant précéder les régions allemandes du symbole A* et les régions italiennes du symbole I*.

La liste des variables construites, et des caractères qui les représentes figure ci-dessous :

1/ Pain, céréales (PAIN)
2/ Viandes, poissons (VIAND)
3/ Produits laitiers, corps gras (LAIT)
4/ Fruits, légumes, pommes de terre (FRUIT)
5/ Produits d'épicerie (EPICE)
6/ Boissons, tabac, repas pris à l'extérieur (BOISS)
7/ Vêtement, chaussures hommes (VETEM)
8/ Vêtement, chaussures femmes (VETEF)
9/ Réparations habillement, soins, santé (SOINS)
10/ Loyers et charges (LOYER)
11/ Energie (ENERG)
12/ Ameublement, équipement (MEUBL)
13/ Textiles d'ameublement, articles ménagers, entretien (MENAG)
14/ Transports (TRANS)
15/ Education, loisirs et divers (EDUCA)

Ce tableau étant homogène (l'addition des éléments d'une ligne ou d'une colonne ayant un sens), nous désirons en obtenir une représentation par l'analyse factorielle des correspondances.

Si le tableau des données s'appelle TAB, il suffit de frapper (cf. annexe III) :

2 *AFCOR TAB*

L'impression des résultats numériques commence aussitôt :

TRACE DE LA MATRICE S 0.04181814

	VALEUR PROPRE	*POURCENTAGE DE LA TRACE*
1	0.016	38.8
2	0.011	27.3

VALEURS DES FACTEURS FI

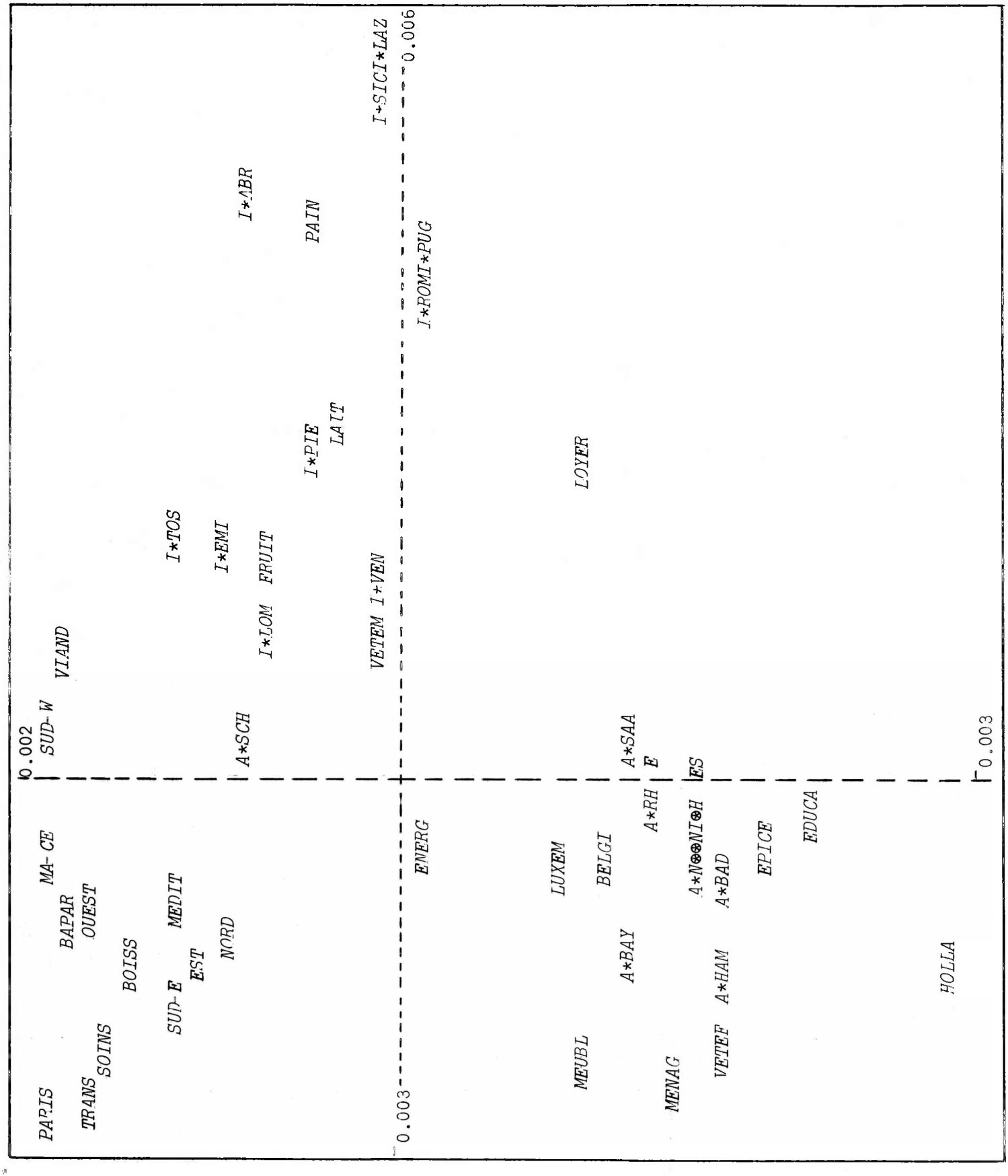
... etc...

Si l'on désire visualiser le plan des deux premiers facteurs, on appelle : (DES désignant le tableau des identificateurs, cf. Annexe VI)

(&FI,FJ) PLANF1F2 DES

La figure 2 est une reproduction du graphique qui apparait alors sur l'imprimante.

EXEMPLE DE SORTIE GRAPHIQUE SUR TERMINAL



Pour obtenir une analyse de la matrice des corrélations des 15 variables, on appelle maintenant : (cf. listage IV)

2 *ACOMP TAB*

On obtient le tableau de résultats numériques reproduit sur la figure 3. On pourrait, bien sur, obtenir une représentation graphique analogue à la précédente.

Pour obtenir un classement des variables selon les valeurs croissantes du premier facteur, il suffit d'indexer le tableau d'identificateurs par les indices rangés à l'aide de l'opérateur ⍋ ; il en est de même si nous désirons le classement des régions selon le premier axe factoriel (les identificateurs des régions sont décalés de 15 places).

Les instructions et les résultats obtenus figurent ci-dessous. On remarque que de telles éditions complémentaires facilitent grandement l'interprétation des directions principales des nuages analysés.

On peut tout aussi facilement, obtenir des histrogrammes, recommencer l'analyse en supprimant un pays ou un groupe de produits ;

DES[⍋*FI*[1;];]

```
TRANS
SOINS
MEUBL
ENERG
VETEF
MENAG
BOISS
VETEM
VIAND
FRUIT
EDUCA
EPICE
LOYER
LAIT
PAIN
```

DES[15+⍋*FJ*[1;];]

```
PARIS
LUXEM
EST
NORD
SUD-E
BELGI
MEDIT
BAPAR
HOLLA
A*HAM
MA-CE
A*NOE
A*SAA
SUD-W
OUEST
I*LOM
A*BAD
A*HES
A*RHE
I*ROM
A*BAY
A*NIE
A*SCH
I*PIE
I*VEN
I*TOS
I*EMI
I*PUG
I*SIC
I*LAZ
I*ABR
```

EXEMPLE D'EDITION DE RESULTATS SUR TERMINAL

2 *ACOMP OUV*

MATRICE DE CORRELATION

	PAIN	*VIAND*	*LAIT*	*FRUIT*	*EPICE*	*BOISS*	*VETEM*	*VETEF*	*SOINS*	*LOYER*	*ENERG*	*MEUBL*	*MENAG*	*TRANS*	*EDUCA*
PAIN	1.000	0.036	0.084	0.094	-0.608	-0.292	-0.143	-0.563	-0.257	0.003	-0.396	-0.571	-0.668	-0.281	-0.452
VIAND	0.036	1.000	0.407	0.678	-0.203	0.780	0.581	0.257	0.784	0.252	0.520	0.335	0.206	0.774	0.096
LAIT	0.084	0.407	1.000	0.427	-0.121	0.149	0.477	0.010	0.193	0.188	0.390	-0.036	-0.231	0.190	-0.178
FRUIT	0.094	0.678	0.427	1.000	0.021	0.576	0.558	0.316	0.664	0.431	0.554	0.290	0.246	0.680	0.083
EPICE	-0.608	-0.203	-0.121	0.021	1.000	0.185	0.188	0.745	0.223	0.116	0.481	0.671	0.776	0.285	0.418
BOISS	-0.292	0.780	0.149	0.576	0.185	1.000	0.430	0.467	0.895	0.087	0.601	0.609	0.580	0.854	0.354
VETEM	-0.143	0.581	0.477	0.558	0.188	0.430	1.000	0.686	0.592	0.406	0.687	0.503	0.397	0.604	0.322
VETEF	-0.563	0.257	0.010	0.316	0.745	0.467	0.686	1.000	0.545	0.360	0.652	0.829	0.858	0.596	0.585
SOINS	-0.257	0.784	0.193	0.664	0.223	0.895	0.592	0.545	1.000	0.106	0.756	0.711	0.629	0.957	0.360
LOYER	0.003	0.252	0.188	0.431	0.116	0.087	0.406	0.360	0.106	1.000	0.286	0.171	0.145	0.194	0.285
ENERG	-0.396	0.520	0.390	0.554	0.481	0.601	0.687	0.652	0.756	0.286	1.000	0.685	0.625	0.710	0.400
MEUBL	-0.571	0.335	-0.036	0.290	0.671	0.609	0.503	0.829	0.711	0.171	0.685	1.000	0.871	0.723	0.578
MENAG	-0.668	0.206	-0.231	0.246	0.776	0.580	0.397	0.858	0.629	0.145	0.625	0.871	1.000	0.639	0.617
TRANS	-0.281	0.774	0.190	0.680	0.285	0.854	0.604	0.596	0.957	0.194	0.710	0.723	0.639	1.000	0.319
EDUCA	-0.452	0.096	-0.178	0.083	0.418	0.354	0.322	0.585	0.360	0.285	0.400	0.578	0.617	0.319	1.000

TRACE DE LA M'TRICE S 15.00000000

	VALEUR PROPRE	*POURCENTAGE DE LA TRACE*
1	7.390	49.3
2	2.963	19.8

VALEURS DES FACTEURS-VARIABLES

FACTEUR NO 1

0.4931 -0.6357 -0.2037 -0.6142 -0.5358 -0.8023 -0.7244 -0.8378 -0.8868 -0.3328 -0.8544 -0.8656 -0.8236 -0.8936 -0.5602

FACTEUR NO 2

0.6007 0.6644 0.6077 0.5865 -0.6781 0.2431 0.2936 -0.3628 0.2696 0.1620 0.0862 -0.3342 -0.5025 0.2488 -0.4224

VALEURS DES FACTEURS-OBSERVATIONS

FACTEUR NO 1

0.3431 -0.2861 0.1794 -0.1565 0.1136 0.1146 0.0789 0.1623 -0.0891 -0.3476 -1.6078 -0.5792 -0.8172 -0.8506 -0.0235
-0.1588 -0.0885 -0.7823 -0.5976 -0.7524 0.3604 0.0632 0.4852 0.6509 0.5868 1.3409 1.3536 1.1104 1.2455
0.1242 -1.1757

FACTEUR NO 2

-0.3284 -0.7282 -0.6400 -0.5706 -0.5543 -0.5386 -0.6877 -0.7295 -0.2758 -0.8141 0.5586 0.3935 0.2339 0.3066 0.0995
0.3248 0.5685 0.1996 0.3015 -0.1435 0.3729 0.3624 0.0681 0.1514 0.2594 0.2899 0.1863 0.2429 0.3270
0.7440 0.0197

REMARQUES :

On peut faire plusieurs remarques à la suite de l'exemple ci-dessus.

1°/ Le terminal pouvait sembler être un goulot d'étranglement de l'information. En fait, les volumineux listages de résultats que l'on obtient lors des exploitations traditionnelles n'étaient justifiés que parce qu'ils fallait tout prévoir à l'avance. En mode conversationnel, on édite seulement ce dont on a besoin, au fur et à mesure de ces besoins.

2°/ La contrainte qui sera le plus rapidement ressentie par le statisticien sera vraisemblablement la taille des espaces de travail. Dans les meilleures installations actuelles, on dispose de 32 K octets. Ceci permet de faire des calculs très importants, car l'occupation de cet espace est gérée "en temps réel" et il est toujours possible de translater ce qui n'est pas immédiatement utile sur d'autres espaces de travail. Cependant, certaines opérations relativement élémentaires, telles que les diagonalisations de matrice, sont grandement facilitées par la mobilisation d'un gros volume de mémoire rapide. Il faut donc souhaiter que les installateurs mettent à la disposition des utilisateurs des zones-mémoires plus importantes.

3°/ Pour les dépouillements d'enquêtes, il sera exclu d'introduire les données à partir du terminal, dans l'état actuel de la diffusion de ces terminaux. Les bandes magnétiques peuvent être domiciliées directement à l'ordinateur, puis copiées sur disques sous forme de fichiers directement interprétables et analysables à partir des terminaux.

4°/ Le contrôle continu du processus de travail donne la possibilité de comparer immédiatement deux méthodes ou techniques, de les combiner éventuellement, enfin de les critiquer. De nombreuses conjectures seront suscitées par ces manipulations.

LISTAGES DE PROGRAMMES D'ANALYSE DE DONNEES USUELS

Nous donnons ci-dessous des exemples de listages de programmes d'un emploi courant en analyse des données. On pourra noter que l'on obtient une édition relativement soignée, avec des instructions réduites et simples. Par suite de la règle de priorité des opérateurs, et de la possibilité de faire plusieurs affectations (symbôle ←) dans une même ligne de calcul, on aura intérêt à lire les lignes de la droite vers la gauche. Par exemple, le symbôle ⍉ désignant la transposition de matrice, la ligne suivante :

```
M ← S+.xT ← ⍉S
```

peut se lire : je mets la transposée de la matrice S dans T que je prémultiplie matriciellement par S, le produit obtenu étant mis dans M (ou étant baptisé M).

Il existe un nombre extrêmement grand de versions A P L d'un même algorithme mathématique. Le caractère synthétique du langage et la possibilité d'affectations répétées dans une même instruction permet d'écrire des versions très condensées de certains programmes, qui sont surtout des exercices de style. Nous préfèrerons une écriture progressive, plus facile à lire ; de toute façon, il s'agit d'exemples qui ne prétendent pas à l'optimalité.

I - Exemple de fonction destinée à faciliter l'introduction de données

```
    ∇ ENTREE
[1]   →(15 1 =ρI←,□)/ 4 0
[2]   'VOUS VOUS ETES TROMPE!'
[3]   →1
[4]   TAB←TAB,[1] I
[5]   →1
    ∇
```

Commentaires : Cette fonction est destinée à permettre un chargement pratique, par le clavier du terminal, du tableau de dimensions 15 x 31 de l'exemple traité plus haut.

Les données sont introduites par séries de 15 (les 15 consommations relatives à une région donnée). L'instruction 1 est un branchement multiple : la parenthèse sera un vecteur à deux composantes prenant les valeurs (1 0) si 15 nombres ont été frappés, (0 1) si un seul nombre a été frappé, (0 0) si le nombre de nombres frappés est différent de 15 ou de 1. En effet, le vecteur des données frappées, □ est mis dans I dont on prend la dimension ρI que l'on compare logiquement au vecteur (15 1). La flèche → signifie "aller à" et le symbôle / utilisé maintenant pour désigner un opérateur dyadique a pour fonction de contracter le vecteur situé à sa droite, selon le vecteur logique de

même dimension situé à sa gauche.

On ira donc en 4 si on a bien frappé 15 données, en 0, c'est-à-dire en fin d'éxécution, si on a frappé une seule donnée, et en séquence si on s'est trompé, ce que notre fonction nous signale en clair. En 4, on construit le tableau TAB de proche en proche, en le complétant à chaque fois du vecteur qui vient d'être lu, et dont la longueur vient d'être contrôlée.

Le volume du commentaire peut paraître important, par rapport au volume et à l'intérêt de la fonction : nous serons plus elliptiques par la suite.

II - Exemple de programme d'extraction des plus grandes valeurs propres d'une matrice symétrique.

On trouvera les algorithmes du programme DIAG ci-dessous et de son auxiliaire HIL dans (BENZECRI, réf. 1).

DIAG extrait les plus grandes valeurs propres d'une matrice symétrique en itérant la transformation linéaire associée, et en orthonormant les vecteurs trouvés à chaque pas, à l'aide de HIL.

```
     ∇DIAG[□]∇
     ∇ Z←CV DIAG S;LV;VI
[1]    VI←S[⍳CV;]
[2]    IT←1
[3]   L1:LV←+/IX×VI←(IX←HIL VI)+.×S
[4]    →L1×⍳(MAXIT≥IT←IT+1)∧ANG<⌈/(H-LV*2)÷H←+/VI*2
[5]    Z←LV,HIL VI
     ∇
```

```
     ∇HIL[□]∇
     ∇ Z←HIL VI;V;VX
[1]    V←1
[2]    →L3,⍴Z←VI
[3]   L1:VX←Z[⍳V-1;]+.×Z[V;]
[4]    Z[V;]←Z[V;]-VX+.×Z[⍳V-1;]
[5]   L3:Z[V;]←Z[V;]÷(+/Z V;*2)*0.5
[6]    →L1×(⍴VI)[1]≥V←1+V
     ∇
```

Commentaires : On désire les CV premiers vecteurs propres de la matrice S. MAXIT désigne le nombre maximum d'itérations désirées, et ANG le carré du sinus du plus grand angle de rotation des vecteurs propres au cours de la dernière itération. On peut raisonnablement fixer pour les applications usuelles MAXIT à 50 et ANG à 10^{-6}.

On notera que les variables qui ne figurent pas sur la ligne du titre de la fonction ne sont pas locales (comme elles le seraient en FORTRAN par exemple).

En tête de ligne, "L1:" ou "L3:" permettent d'identifier la ligne, quelle que soit la place de cette ligne dans la fonction, ce qui donne plus de souplesse à l'écriture et au test des programmes.

L'opérateur ⍳ (générateur d'indice), appliqué à un nombre entier K, génère un vecteur dont les K composantes sont 1, 2, 3, K-1,K.

Ainsi, dans DIAG, la matrice VI, qui sert au "démarrage" de l'itération, est formée des CV premières lignes de S.

III - Exemple de programme d'analyse des correspondances

Comme le programme d'analyse en composantes principales que l'on trouvera ci-dessous, AFCOR utilise, outre DIAG, deux petits programmes auxiliaires d'édition : EDIFAC et EDIPROPRE, dont la liste est donnée plus loin.

```
     ∇ Z←NF AFCOR P;S;K;PI;PJ
[1]    PI←+/P←P÷K←+/+/P
[2]    PJ←+⌿P
[3]    S←(PI∘.×PJ)*0.5
[4]    P←(P÷S)-S
[5]    Z←NF DIAG S←S+.×⍉S
[6]    (+/ 1 1 ⍉S) EDIPROPRE Z[;1]
[7]    'VALEURS DES FACTEURS FI'
[8]    ,'----------------------'
[9]    EDIFAC FI←(0 1 ↓Z)×(Z[;1]∘.÷PI)*0.5
[10]   'VALEURS DES FACTEURS FJ'
[11]   ,'----------------------'
[12]   EDIFAC FJ←(FI+.×P)÷(Z[;1]*0.5)∘.×PJ
     ∇
```

Commentaires : L'analyse du tableau de nombres positifs P donne lieu à l'extraction de NF facteurs. On diagonalise la Matrice S "symétrisée", calculée à partir des inerties par rapport au centre de gravité, et non par rapport à l'origine, afin d'éliminer le facteur trivial.

IV - Exemple de programme d'analyse en composantes principales

```
     ∇ Z←NF ACOMP P
[1]    Q←⍴P[;1]
[2]    M←(+/P)÷N←⍴P[1;]
[3]    S←(((+/P*2)÷N)-M*2)*0.5
[4]    P←(P-M∘.×(N⍴1))
[5]    P←P÷S∘.×N⍴1
[6]    Z←NF DIAG COR←(P+.×⍉P)÷N
[7]    'MATRICE DE CORRELATION'
[8]    ,'----------------------'
[9]    ((9⍴' '),⍕2↑,M),[1] ' ',[1](M←1⍕' ',IDENT,' '), 7 3
                                                    DFT COR
[10]   (+/ 1 1 ⍉COR) EDIPROPRE Z[;1]
```

```
[11]  'VALEURS DES FACTEURS VARIABLES'
[12]  ,'----------------------------',
[13]  EDIFAC FI←(0 1 ⍉Z)⍉(Z[;1]∘.×Q⍴1)*0.5
[14]  'VALEURS DES FACTEURS OBSERVATIONS'
[15]  ,'--------------------------------',
[16]  EDIFAC FJ←(FI+.×P)÷(Z[;1]*0.5)∘.×N⍴Q*0.5
      ∇
```

Commentaires : Le tableau initial P contient les variables ou attributs en ligne et les observations ou individus en colonnes. La première instruction désigne par Q le nombre de variables (dimensions d'un vecteur colonne). La seconde désigne par N le nombre d'observations, et calcule les moyennes relatives à chacune des variables. Le vecteur des écarts-types est calculé en 3, le tableau P est centré en 4, réduit en 5; la matrice des corrélations COR, calculée et diagonalisée en 6, est imprimée en 9. La fonction DFT, disponible sur la plupart des installations A P L, permet de fixer le nombre de décimales à l'impression. Pour plus de détail concernant le cadrage variable-observation, on pourra se reporter à réf. 5

V - Exemples de programmes d'édition de résultats en analyse des données.

```
      ∇ I EDIPROPRE J;K
[1]   2⍴RC;'TRACE DE LA MATRICE S'; 20 8 DFT I
[2]   K←1
[3]   '        VALEUR PROPRE      POURCENTAGE DE LA TRACE'
[4]  L1:K; 15 3 20 1 DFT J[K],100×J[K]÷I
[5]   →L1×⍳(⍴J)≥K←K+1
      ∇
```

```
      ∇ EDIFAC I;K
[1]   K←1
[2]  L1:'FACTEUR NO ';K
[3]   8 4 DFT I[K;]
[4]   →L1×⍳(⍴I)[1]≥K←K+1
      ∇
```

Commentaires : Ces deux petites fonctions, appelées par AFCOR et par ACOMP, utilisent également la fonction d'édition DFT. Pour la présentation des résultats, on pourra se reporter à l'exemple traité plus haut.

VI - Exemples de programmes de représentation graphique sur machine à écrire du terminal.

La fonction PLANF1F2, qui appelle la fonction FUSION , a servi à tracer le graphique de l'exemple ci-dessus. Elle utilise une autre fonction auxiliaire, D4 I, qui n'est autre que 6 3 DFT I, c'est-à-dire une écriture du tableau I avec des nombres de 6 caractères, dont 3 après la virgule.

```
     ∇ TT←XY PLANF1F2 DES;IO;IA;SO;SA;J;T;TT;K
[1]    IO←⌊/XY[;2]
[2]    SO←⌈/XY[;2]
[3]    IA←⌊/XY[;1]
[4]    SA←⌈/XY[;1]
[5]    J←⌈1+(XY[;2]-IO)×(L-1)÷SO-IO
[6]    T←(-⌈1+(XY[;1]-IA)×(C-1)÷SA-IA)⌽DES,((1↑ρDES),C)ρ' '
[7]    TT←(L,(ρT)[2])ρ' '
[8]    I←L
[9]   L1:→(1 0 =ρK←(J=I)/⍳ρJ)/L2,L4
[10]   TT[I;]←FUSION T[K;]
[11]   →L4
[12]  L2:TT[I;]←T[K;]
[13]  L4:→(⍳I≤I←I-1)/L1
[14]   T←⍉0
[15]   TT←⊖TT
[16]   TT←(-J)⌽[1]((D4 IA),(((ρTT)[2]-12)ρ'-'),D4 SA),[1](J←⌈L×SO÷SO-IO)⌽[1] TT
[17]   J←⌈C×-IA÷SA-IA
[18]   TT←(-J)⌽(((ρTT)[2]+1)↑D4 SO,[1]('|',J⌽TT),[1]((ρTT)[2]+1)D4 IO
[19]   TT←'_',[1]('|',TT,'|'),[1]'_'
     ∇
```

```
     ∇ Z←FUSION M;I;J;D
[1]    Z←(D←ρM)[2]ρ' '
[2]    J←D[2]|I←(M≠' ')/⍳ρM←,M
[3]    Z[J+D[2]×0=J]←M[I]
[4]    Z[J[(1≠+/J∘.=J)/⍳ρJ]]←'⊛'
     ∇
```

Commentaires : Le tableau DES contient les identificateurs des points (par exemple, cinq lettres par point, comme dans notre exemple). S'il y a N points, DES a alors pour dimension (N,5). Les deux colonnes du tableau XY sont les valeurs numériques des abscisses et des ordonnées des points du nuage à représenter. Les paramètres L et C, qui ne sont pas des variables locales, sont les nombres de lignes et de colonnes du graphique désiré. La fonction FUSION sert à marquer d'un symbôle particulier les points multiples, repérés au cours de l'instruction 9.

REFERENCES BIBLIOGRAPHIQUES DE L'ANNEXE

1 - BENZECRI J.P. (1970) - Algorithmes de géométrie euclidienne en analyse des données - (in "l'Analyse des données" - Tome 2 - DUNOD 1973).

2 - A Programming language - K.E. IVERSON - John WILEY and Son - New-York - 1962

3 - L'informatique par téléphone : une introduction au langage A P L - P.S. ABRAMS et G. LACOURLY - Hermann 1972.

2 Ex - n°. 1