Statistique et analyse des données

M. Depaix

Exemples d'interprétations multifactorielles

Statistique et analyse des données, tome 1, n° 2 (1976), p. 104-115

<http://www.numdam.org/item?id=SAD_1976__1_2_104_0>

Exemples d'interprétations multifactorielles

M. DEPAIX

L'article qui suit est extrait d'une conférence donnée au Congrès annuel de la Société d'Hygiène, de Médecine Sociale et de Génie Sanitaire. Il s'agissait de présenter un certain nombre d'études utilisant des méthodes d'analyse des données. Je n'ai retenu ici que deux études pour ne pas alourdir le texte, renvoyant pour les autres à la revue de la Société d'hygiène où le texte complet paraîtra.

Les différentes méthodes statistiques élaborées en liaison avec l'admirable outil que constitue un ordinateur permettent maintenant de lancer des études portant sur un nombre important de caractères d'un même individu.

L'accumulation de données duement contrôlées qui, à première vue, parait être un facteur de progrès dans la connaissance, devient vite un obstacle à ce même progrès du fait que l'homme n'est pas capable d'assimiler puis d'interpréter un grand nombre de données.

La nécessité d'extraire de vastes tableaux de nombres les éléments prépondérants s'est vite fait sentir et a conduit soit à remettre au premier plan des méthodes statistiques élaborées depuis la fin du 19e siècle (analyse en composantes principales, analyse factorielle) soit à en élaborer de nouvelles (analyse des correspondances). Ces différentes analyses procèdent en général de la même idée : construire un nuage de points représentant l'ensemble des données puis le projeter dans un espace de faible dimension en respectant au maximum la "forme du nuage". Cette représentation permet alors soit de dégager les premiers éléments d'une classification soit de mettre en évidence des groupements de caractère.

Nous présenterons dans la suite quelques résultats d'analyses en composantes principales et d'analyses factorielles des correspondances pour illustrer ce qui vient d'être dit....

EXEMPLES

1°) Exemples d'analyses factorielles des correspondances

a) Etude d'un profil socio-professionnel et socio-culturel (figure 1)

On a utilisé un échantillon de 1673 hommes de 40 à 50 ans. L'objectif principal était l'élaboration de profils socio-professionnels et socio-culturels à l'aide de données recueillies par questionnaire lors du bilan de santé, l'objectif secondaire étant la réduction du questionnaire à quelques paramètres clefs.

Sur les 47 questions du questionnaire, 12 ont été retenues (les autres étant presque inutilisables par exemple du fait de la trop forte concentration des réponses sur une seule modalité) totalisant 59 modalités.

L'analyse factorielle des correspondances est un outil bien adapté à ce genre de recherche. Le profil d'un individu pourra être constitué par l'ensemble de ses coordonnées sur les premiers axes factoriels.

Dans l'exemple présenté l'inertie du 1er axe représente 7,6 % de l'inertie totale, celle du second en étant les 4,7 %. Ce sont donc de faibles inerties mais finalement ces axes sont importants du point de vue discrimination. Un effet Gutman apparait : on pourrait penser que seul le 1er axe est intéressant, mais des interprétations ont été données pour les 3 premiers axes.

Le 1er facteur semble représenter le statut socio-professionnel. Il ordonne parallèlement les catégories socio-professionnelles (CSP), les vacances et les diplômes, et de plus les 4 caractères : CSP de l'individu, de son père, diplôme et vacances expliquent plus de 70 % du 1er facteur.

Le 2e axe est susceptible de 2 interprétations : en premier lieu l'interprétation classique associée à l'effet Gutman consiste à dire que le 2é facteur mesure l'intensité du premier ; en deuxième lieu, constatant que le 2e axe oppose

jamais de télévision à beaucoup de télévision

maison individuelle à immeuble

hommes 40-50 ans

FIGURE 1

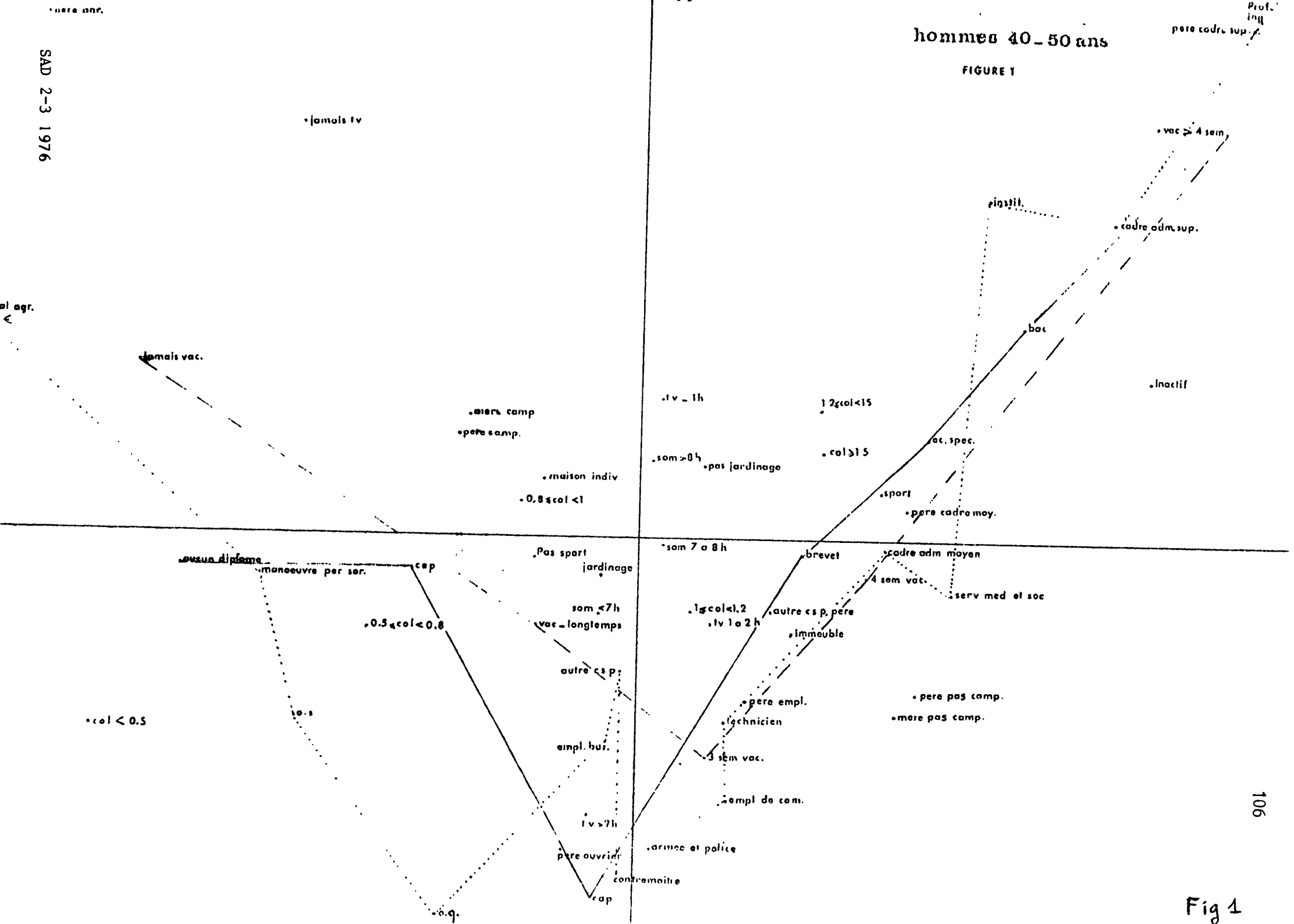


Fig 1

grand logement à petit logement

professions assez indépendantes

(agriculteurs, professions libérales, professeurs, ingénieurs)

à

professions plus dépendantes sur le plan socio-économique

(employés, contremaîtres, ouvriers)

On est amené à suggérer qu'il reflète un facteur de dépendance socio-économique. En combinant les deux interprétations on peut dire que la dépendance est parallèle à l'intensité du statut socio-professionnel :

à une forte intensité correspond une grande indépendance

à une faible intensité correspond une grande dépendance

b) Evolution avec l'âge du profil socio-professionnel et socio-culturel (figures 2α, 2β, 2γ, 2δ)

Une objection à l'étude précédente pourrait être une possible instabilité avec l'âge des résultats. Pour se libérer de ce doute on a effectué plusieurs études suivant différentes tranches d'âge. On ne présente ici que quelques exemples

Figure 2α

L'étude porte sur un échantillon de 2188 hommes de 20 à 30 ans

L'inertie du 1er axe est de 6,1 % celle du 2e axe est de 4,6 %.

Figure 2β

L'étude porte sur un échantillon de 2178 hommes de 30 à 40 ans

L'inertie du 1er axe est de 6,1 % celle du 2e axe est de 4,6 %.

Figure 2γ

L'étude porte sur un échantillon de 1816 hommes de 40 à 50 ans

L'inertie du 1er axe est de 6,6 % celle du 2e axe est de 4,6 %.

Figure 2 δ

L'étude porte sur un échantillon de 1034 hommes de 50 à 60 ans

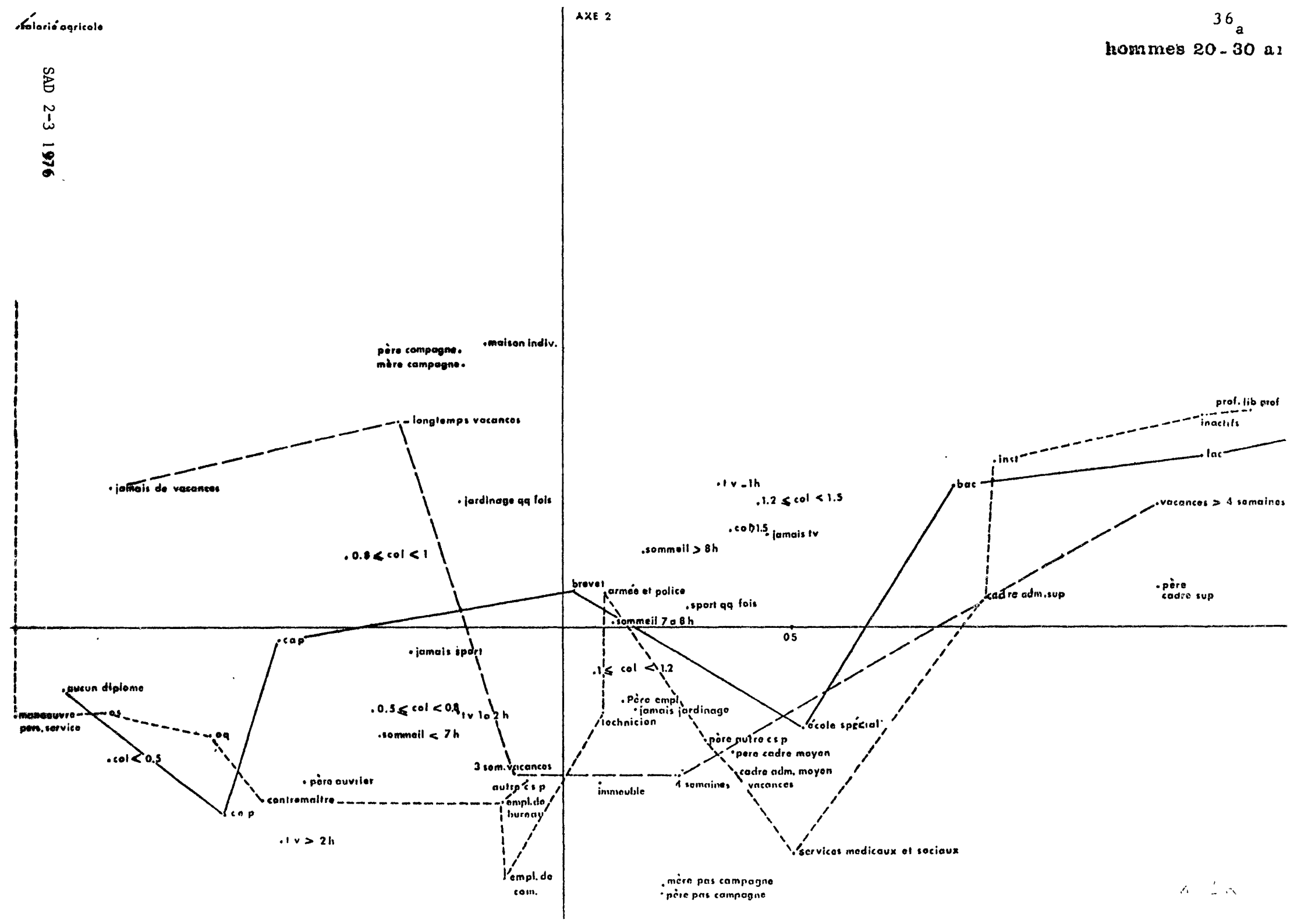
L'inertie du 1er axe est de 6,6 % celle du 2e axe est de 4,3 %.

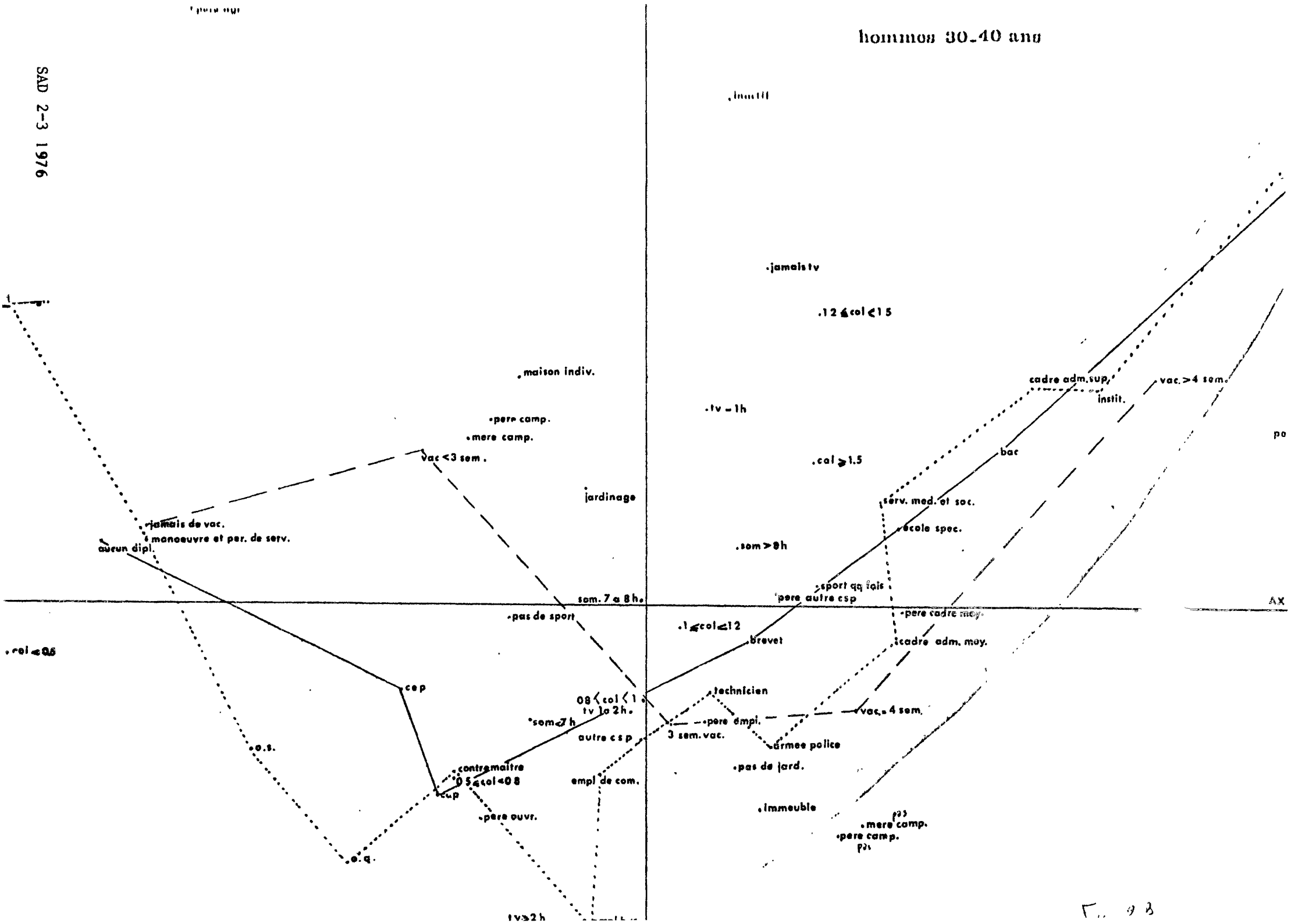
On remarque déjà la stabilité de l'inertie des 2 premiers axes. Puis on note que les schémas β, γ δ sont très semblables à ce qui a été obtenu dans la première étude. Seul α est légèrement différent.

hommes 20 - 30 a 1

salarié agricole

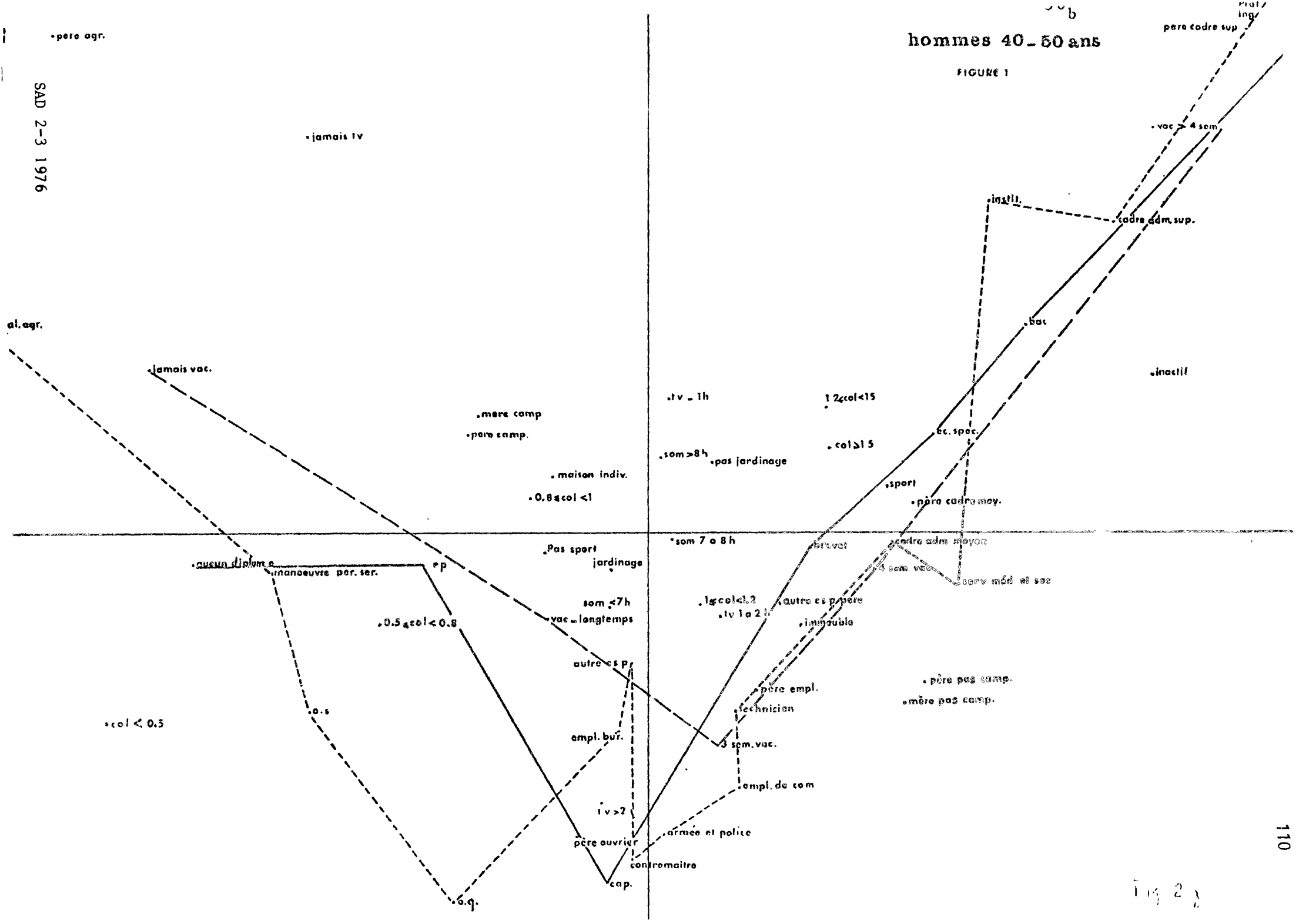
SAD 2-3 1976




hommes 30-40 ans
SAD 2-3 1976
jamais tv
1.2 ≤ col ≤ 1.5
maison indiv.
cadre adm.sup.
vac.>4 sem.
instit.
tv - 1h
pere camp.
mere camp.
vac <3 sem.
bac
col ≥ 1.5
jardinage
serv. med. et soc.
jamais de vac.
manoeuvre et per. de serv.
école spec.
aucun dipl.
som > 8h
sport qq fois
pere autre csp
AX
som. 7 a 8 h.
pere cadre moy.
pas de sport
1 ≤ col ≤ 1.2
brevet
cadre adm. moy.
col ≤ 0.6
cap
technicien
0.8 < col < 1
tv 1a 2h.
vac.= 4 sem.
som<7 h
pere empl.
3 sem. vac.
autre c s p
armee police
o.s.
contremaitre
0.5 ≤ col < 0.8
pas de jard.
empl de com.
cap
pere ouvr.
immeuble
mere comp.
pere comp.
o.q.
tv≥2 h

hommes 40_50 ans

FIGURE 1



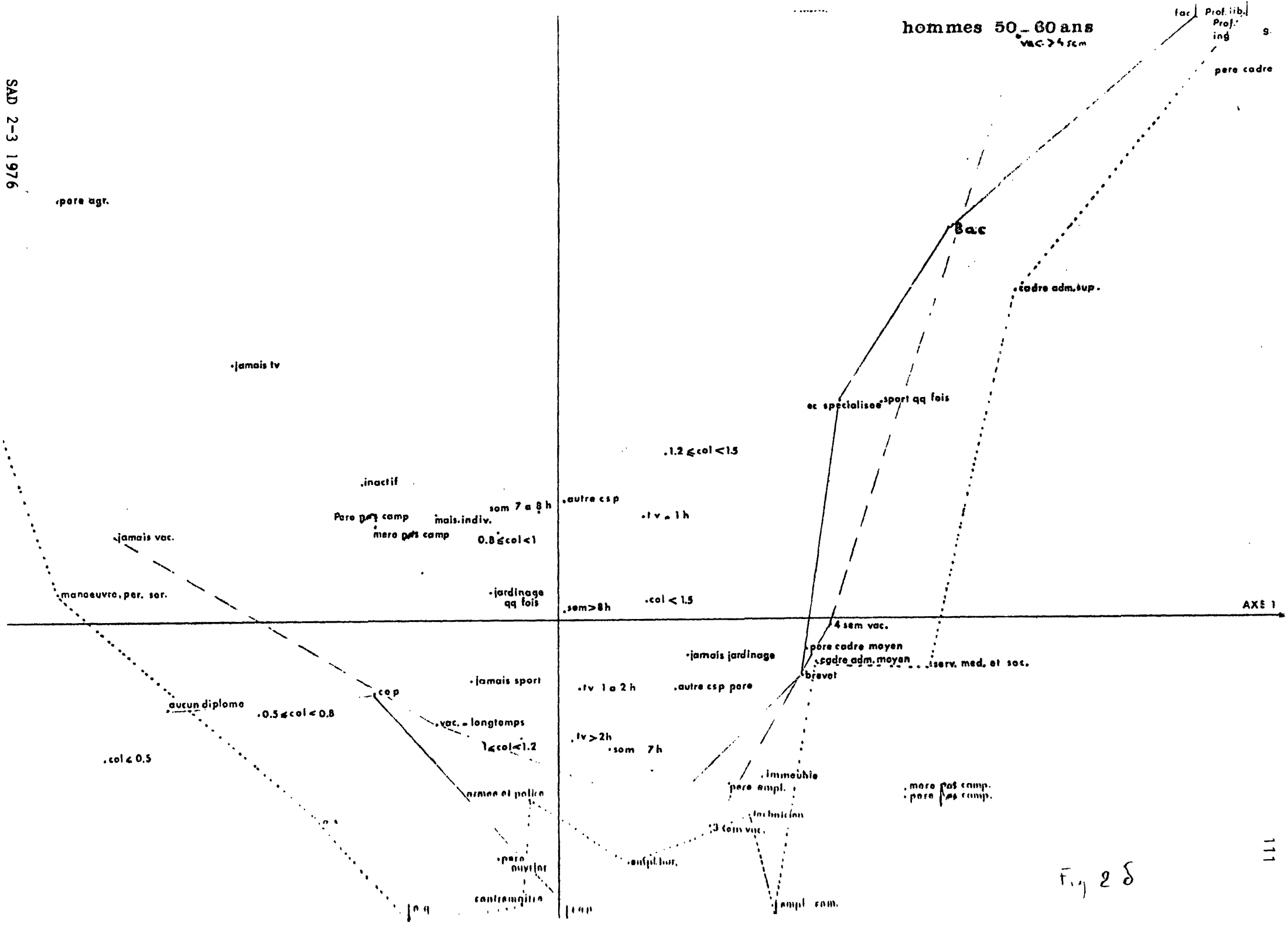

hommes 50 - 60 ans
vac. > 4 sem
fac
Prof. lib.
Prof. ing
pere cadre
Bac
cadre adm. sup.
ec specialisee
sport qq fois
pere agr.
jamais tv
1.2 ≤ col < 1.5
inactif
autre csp
som 7 a 8 h
tv = 1 h
Pere pas camp
mais. indiv.
mere pas camp
jamais vac.
0.8 ≤ col < 1
jardinage qq fois
manoeuvre, per. sor.
sem > 8h
col < 1.5
AXE 1
4 sem vac.
jamais jardinage
pere cadre moyen
cadre adm. moyen
serv. med. et soc.
brevet
jamais sport
cap
tv 1 a 2 h
autre csp pere
aucun diplome
0.5 ≤ col < 0.8
vac. = longtemps
1 ≤ col < 1.2
tv > 2h
som 7h
col < 0.5
immeuble
pere empl.
mere pas camp.
pere pas camp.
armee et police
instituteur
3 sem vac.
empl. bur.
pere ouvrier
contremaitre
ouvr. qual.
empl. com.
cap

Fig 2 δ

Ce qui évolue fortement, ce sont les contributions relatives des caractères aux facteurs : les contributions de la CSP de l'individu et de son diplôme diminuent quand l'âge augmente à l'inverse des contributions de la CSP du père et de l'origine des parents.
Le statut socio-professionnel serait donc davantage déterminé par le diplôme et la profession chez les plus jeunes et par l'origine et la profession des parents chez les plus âgés comme l'indique le tableau suivant qui donne la contribution des caractères de la colonne de gauche au 1er facteur.

TRANCHES D'AGE	20 - 30	30 - 40	40 - 50	50 - 60
CSP de l'individu	26,2 %	23,4 %	22,8 %	21 %
CSP de son père	13,4 %	15 %	17,2 %	16,1 %
Diplôme	24,2 %	20 %	15,2 %	13,9 %
Origine du père	1,7 %	5,4 %	7,7 %	9,8 %
Origine de la mère	1,6 %	6,7 %	6,8 %	9,8 %

Les exemples a et b correspondent à des analyses factorielles des correspondances de type ensembliste ; le tableau des données comporte en colonne les variables (ou plutôt leurs modalités) et en ligne les individus. Chaque case ne contient alors que la valeur 0 ou 1 selon que l'individu considéré ne présente pas ou présente la modalité associée à la case.
L'exemple c correspond à un type différent : les variables colonnes ~~représentent~~ les modalités des paramètres socio-professionnels et les variables lignes sont les modalités des paramètres fonctionnels et biologiques. Chaque case contient le nombre d'individu possédant les 2 modalités.
On a représenté dans le même plan les projections des 2 types de caractères.

c) Etude des liaisons entre caractères socio-professionnels et paramètres fonctionnels et biologiques (figures 3Ø, 36)
L'étude a porté sur 5592 hommes de 20 à 50 ans. Les variables quantitatives ont été découpées en classes ; on a ainsi mis en présence 33 paramètres fonctionnels et biologiques découpés en 146 modalités et caractères socio-professionnels découpés en 37 modalités.

Seuls quelques paramètres sont reportés sur la ligne 3 . Il faut noter que les numéros des 5 classes des paramètres quantitatifs sont dans l'ordre de grandeur de ces paramètres.

On a là un exemple de fort taux d'inertie pour les deux premiers axes : 38,1 et 11,6 %, soit près de 50 % pour les deux réunis.
Le 1er axe semble toujours être un axe de statut professionnel. Mais les liens entre les paramètres biologiques et les paramètres socio-professionnels ne sont pas très clairs.

La figure suivante (3Ø) illustre d'une façon plus p-écise ce qu'on vient de voir : les répartitions des caractères dans deux cas extrêmes (ingénieurs, O.S.) sont nettement différents. On comprend mieux ici les positions des différentes projections les unes par rapport aux autres : c'est, par exemple, l'opposition de forme des distributions de la GoT entre les ingénieurs et cadres administratifs supérieurs d'une part, et entre les OS, les manoeuvres et les salariés agricoles d'autre part, qui conditionne les différentes positions des points GoT sur le diagramme des professions.

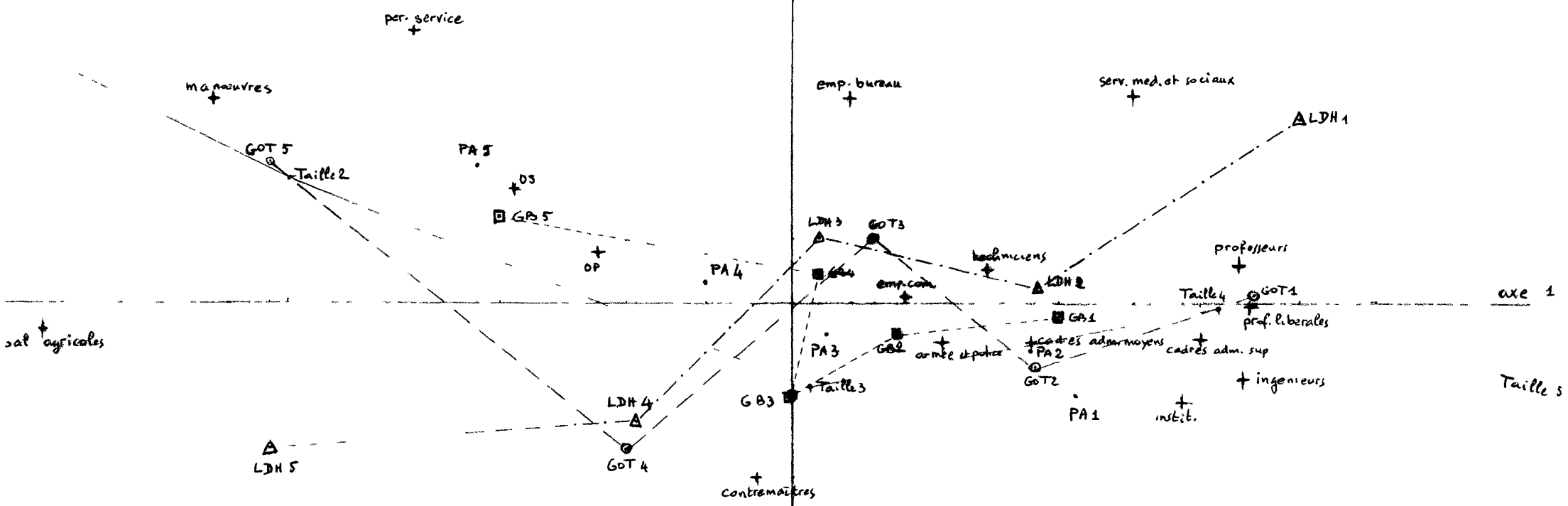
Hommes 20-50 ans

$T_1 = 38,1\%$ $T_2 = 11,6\%$

axe 2

Analyse des correspondances

Param. Socio-Prof – Param. fonct. et biolog



PA = Phosphatases alcalines

GB = Globules blancs

OS = ouvriers specialisés

OP = ouvriers professionnels

Fig 3 a

DIAGRAMMES DE QUELQUES DISTRIBUTIONS DANS DEUX GROUPES SOCIO-PROFESSIONNELS

INGENIEURS et CADRES ADM. SUPERIEURS		OS, MANOEUVRES, SALARIES AGRICOLES