Introduction à l'analyse des données

Olivier Godechot

Introduction.

- Les données statistiques : de très nombreuses variables.
- Aucune n'est parfaite
- La perception d'un phénomène appréhendée comme la combinaison d'un grand nombre de variables
- Comment faire pour tenir compte de l'ensemble de l'information ?

Introduction. Suite

- Faire des tableaux croisés (variable x*y) ou calculer des coefficients de corrélation.
- Problème : si 10 variables ->45 tableaux croisés. Si 100 variables, 4950 tableaux croisés
- Autre méthode : les indices
- Exemple : indice d'inflation. Indice de développement humain, BIP40
- I=a1.X1+a2.X2+a3.X3
- Problème : arbitraire de la formule et des pondérations

Introduction. Fin

- Trouver des méthodes pour synthétiser les variables sans trop les déformer.
- Trouver des axes (qui sont alors des indices) qui respectent la forme du nuage multidimensionnel, c'est-à-dire la forme de la relation entre les variables.

Les différentes méthodes

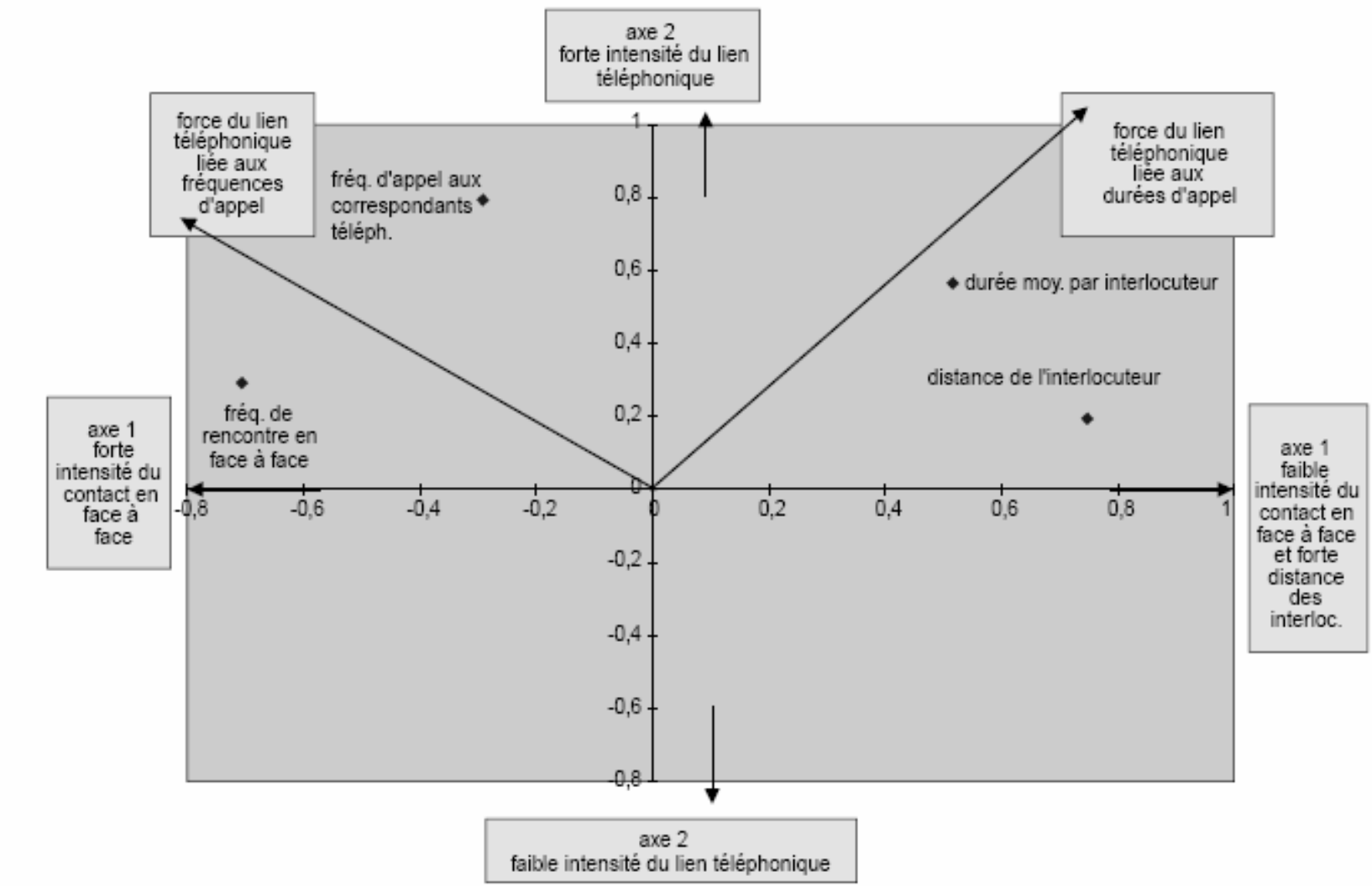
- Les méthodes factorielles de représentation
  - Analyse en composantes principales (ACP). Variables quantitatives.
  - Analyse factorielle de correspondances (AFC). Analyse d'un tableau croisé de deux variables qualitatives.
  - Analyse des correspondances multiples (ACM). Plusieurs variables qualitatives.
- Les méthodes de classification (des individus)
  - Classification ascendante hiérarchique (CAH)
  - Classification autour des centres mobiles
- La discrimination

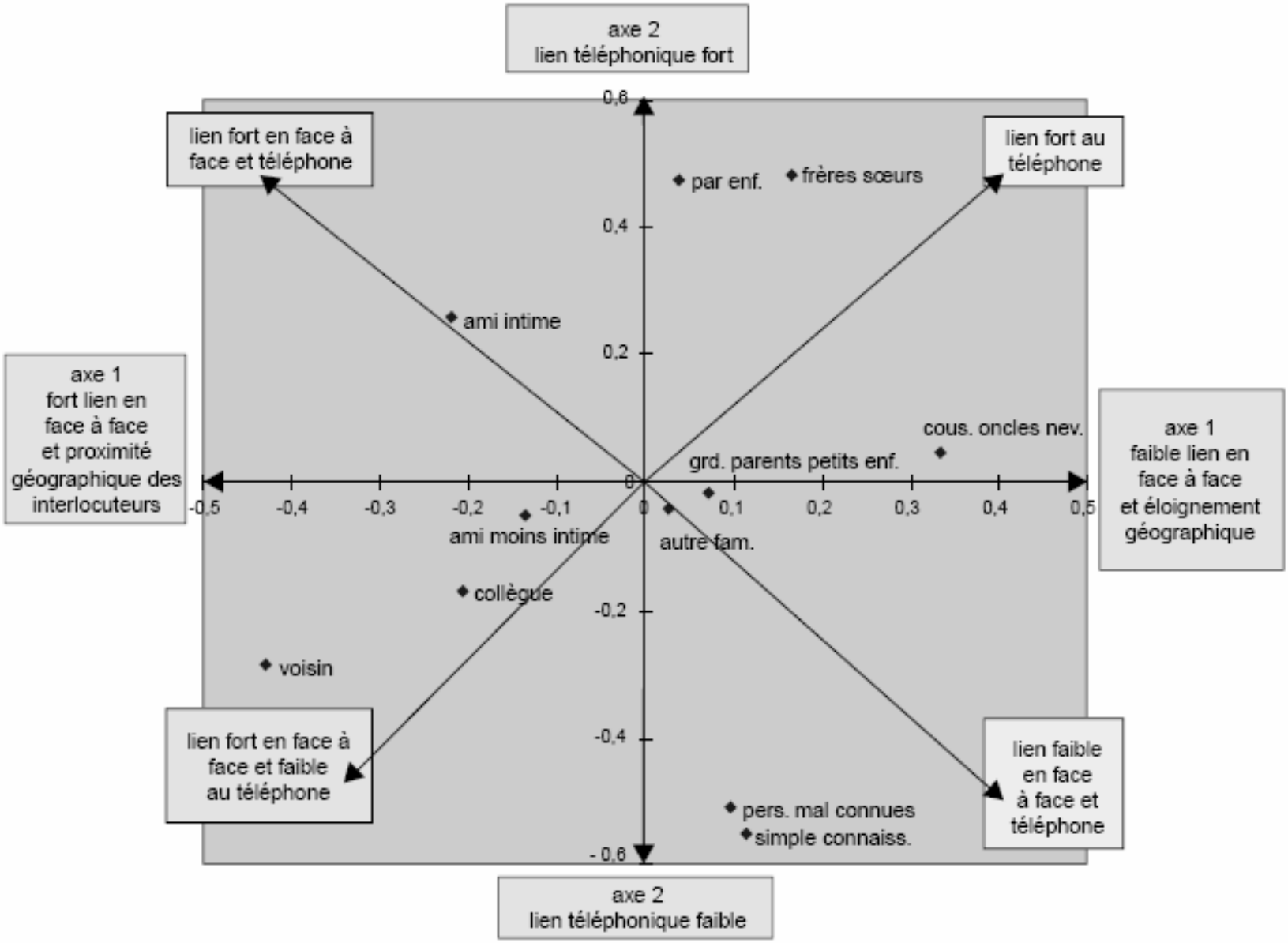
L'analyse en composantes principales

- Historiquement, la plus ancienne (1901, Pearson).
- Utilisation importante par les psychologues du QI.
- Relativement peu utilisée en sociologie
- Plus facile à expliquer
- AFC et ACM sont des ACP particulières

Un exemple de mise en oeuvre

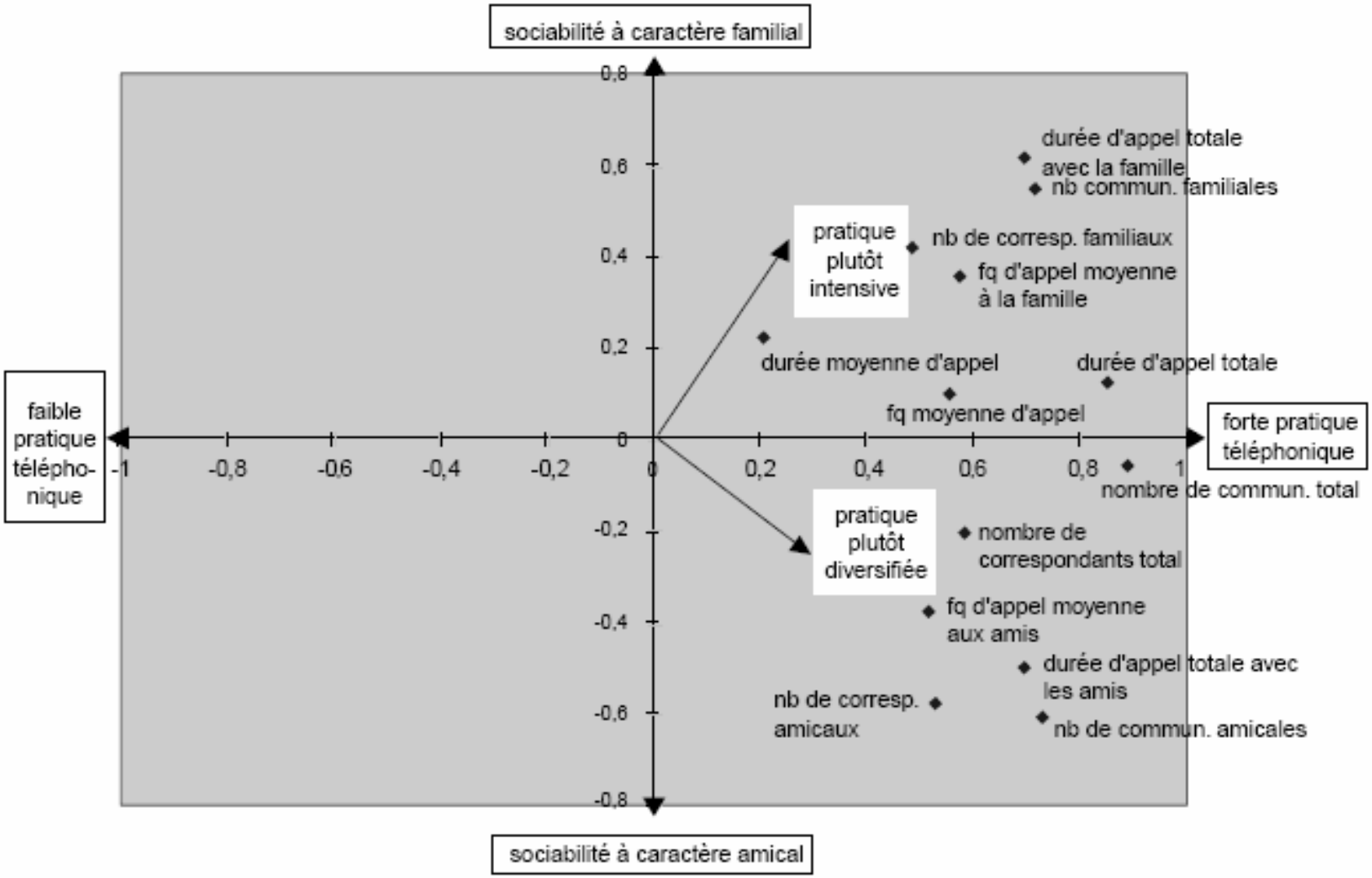
Carole-Anne Rivière, 2001, « Le téléphone un facteur d'intégration sociale », Économie et Statistiques, n°345.




axe 2
lien téléphonique fort
lien fort en face à face et téléphone
lien fort au téléphone
par enf.
frères sœurs
ami intime
axe 1
fort lien en face à face et proximité géographique des interlocuteurs
axe 1
faible lien en face à face et éloignement géographique
cous. oncles nev.
grd. parents petits enf.
ami moins intime
autre fam.
collègue
voisin
lien fort en face à face et faible au téléphone
lien faible en face à face et téléphone
pers. mal connues
simple connaiss.
axe 2
lien téléphonique faible
0,6
0,4
0,2
0
-0,2
-0,4
- 0,6
-0,5
-0,4
-0,3
-0,2
-0,1
0
0,1
0,2
0,3
0,4
0,5

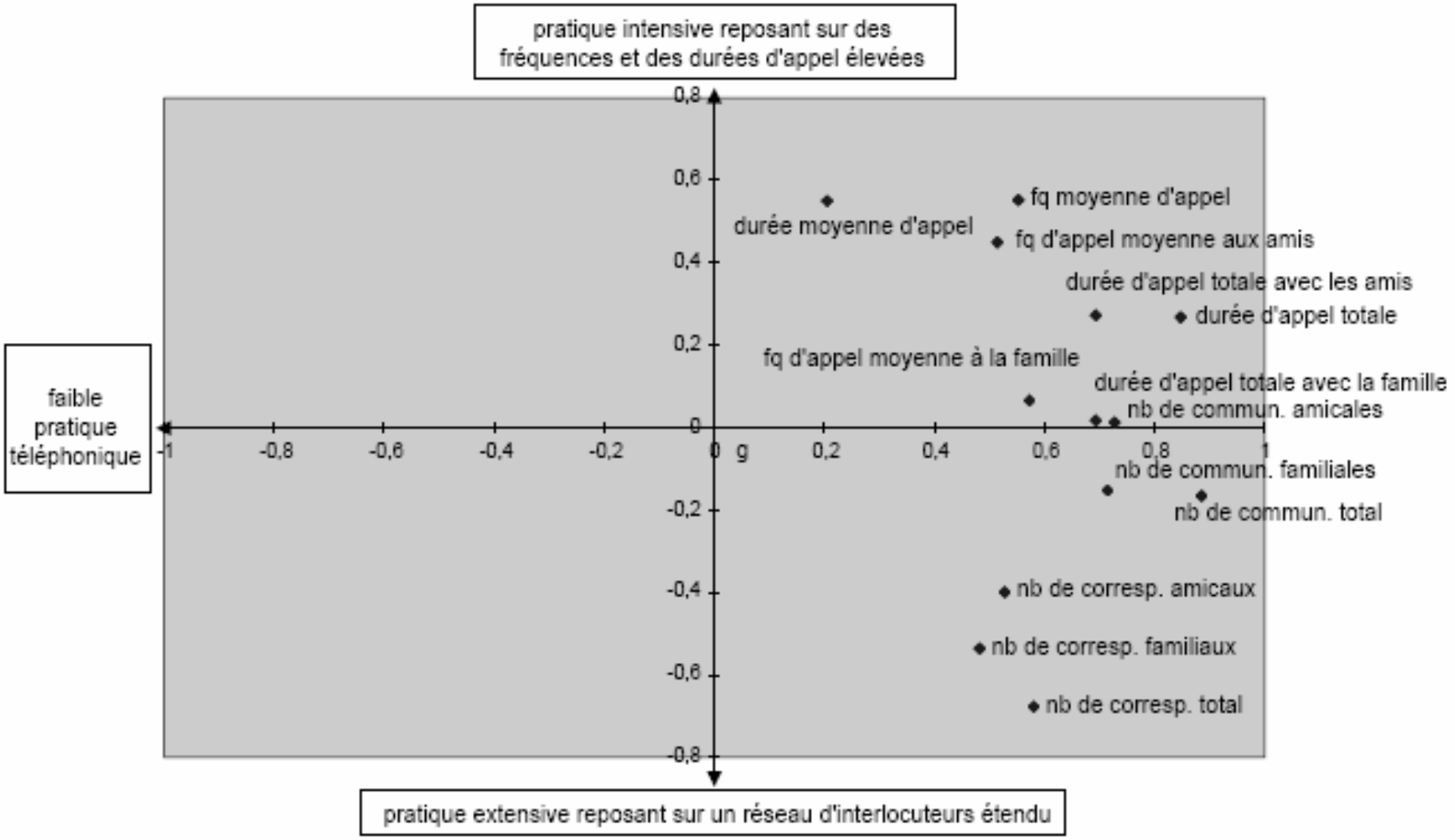
Graphique III

Composante de la sociabillité téléphonique - Axes 1 et 2



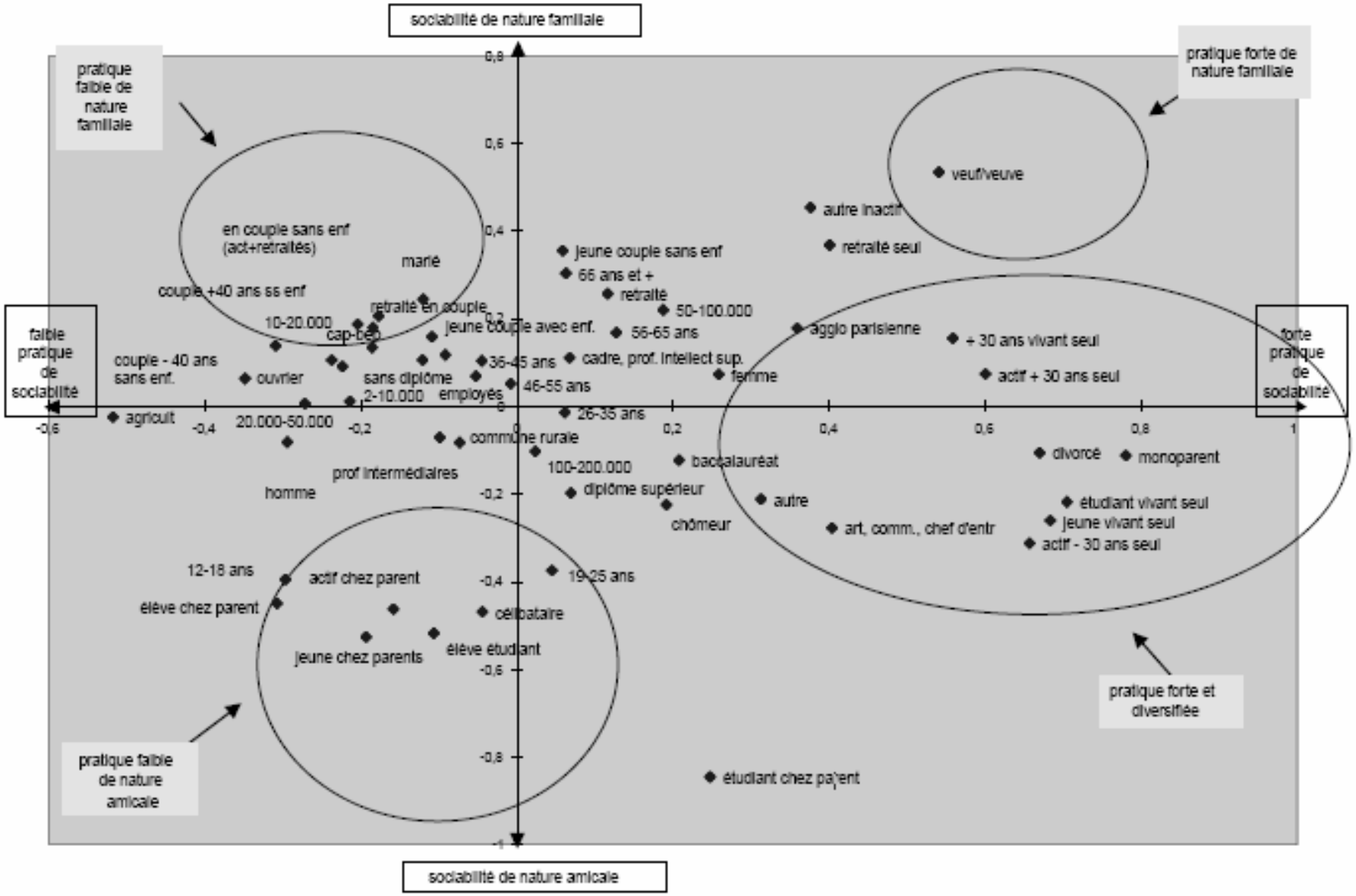
Graphique IV

Composante de la sociabilité téléphonique - Axes 1 et 3



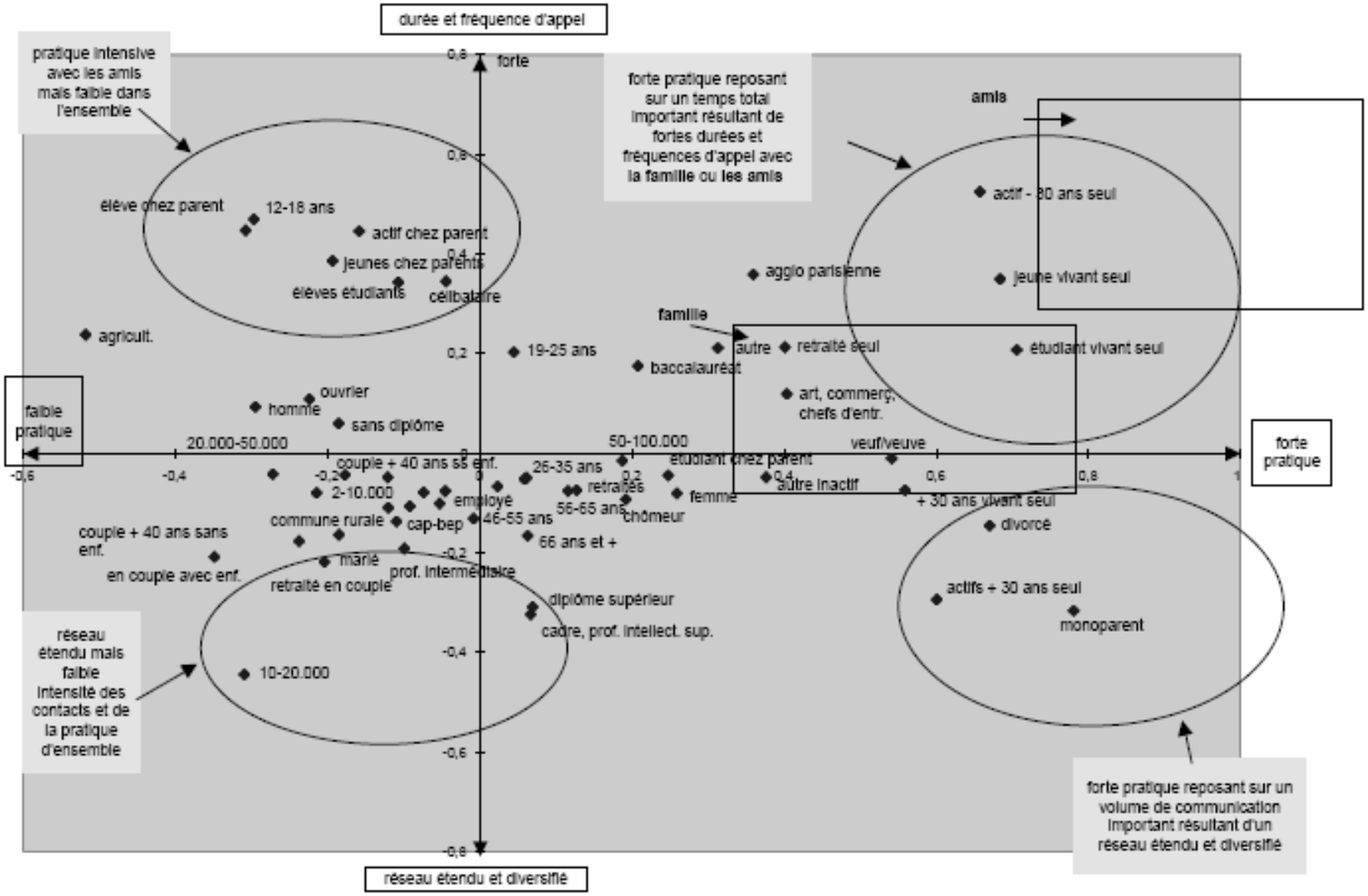
Graphique V

Espace de la sociabilité téléphonique - Plan principal, axes 1 et 2



Graphique VI

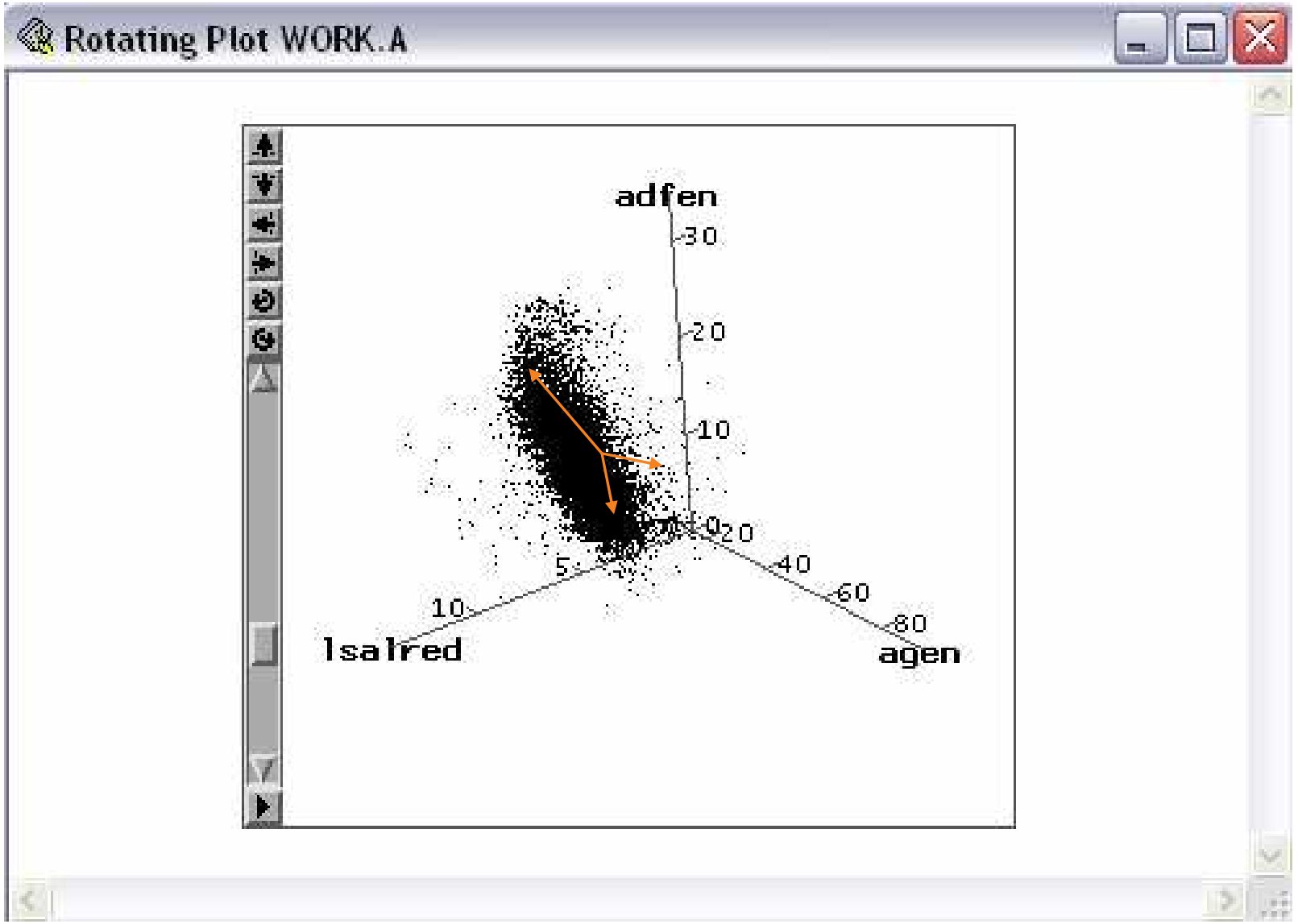
Espace de la sociabilité téléphonique - Plan principal, axes 1 et 3



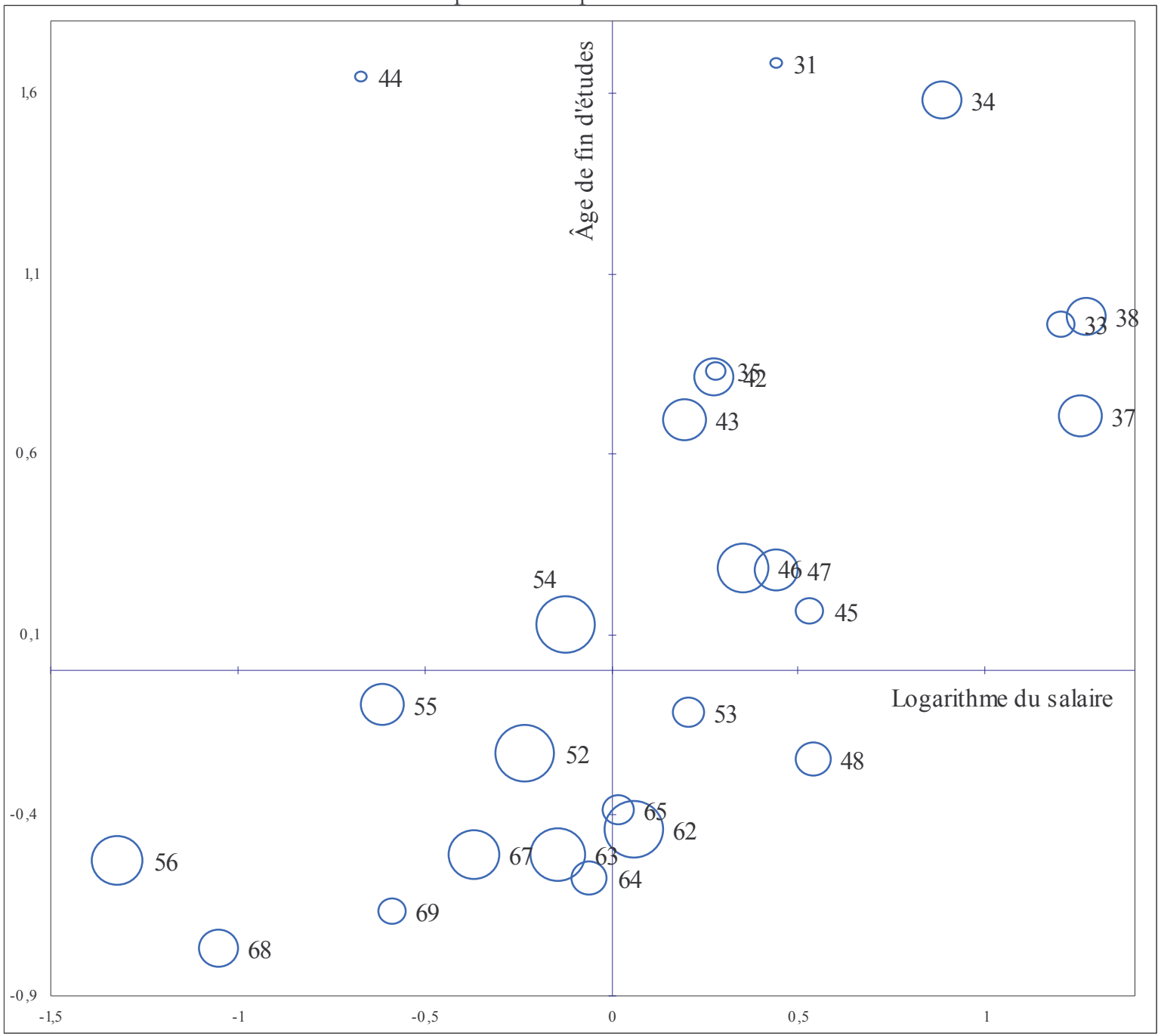
ACP. L'idée

- Représenter des données multidimensionnelles sur un nombre limité de dimensions en déformant au minimum les données
- La méthode : un changement de repères.
- Les axes du nouveau repère sont calculés à partir des axes originaux de l'ancien repère.
- Ces nouveaux axes sont hiérarchisés en fonction du degré de fidélité aux distances entre les points dans le nuage.
- CQ : les premiers axes donnent une assez bonne représentation des données.

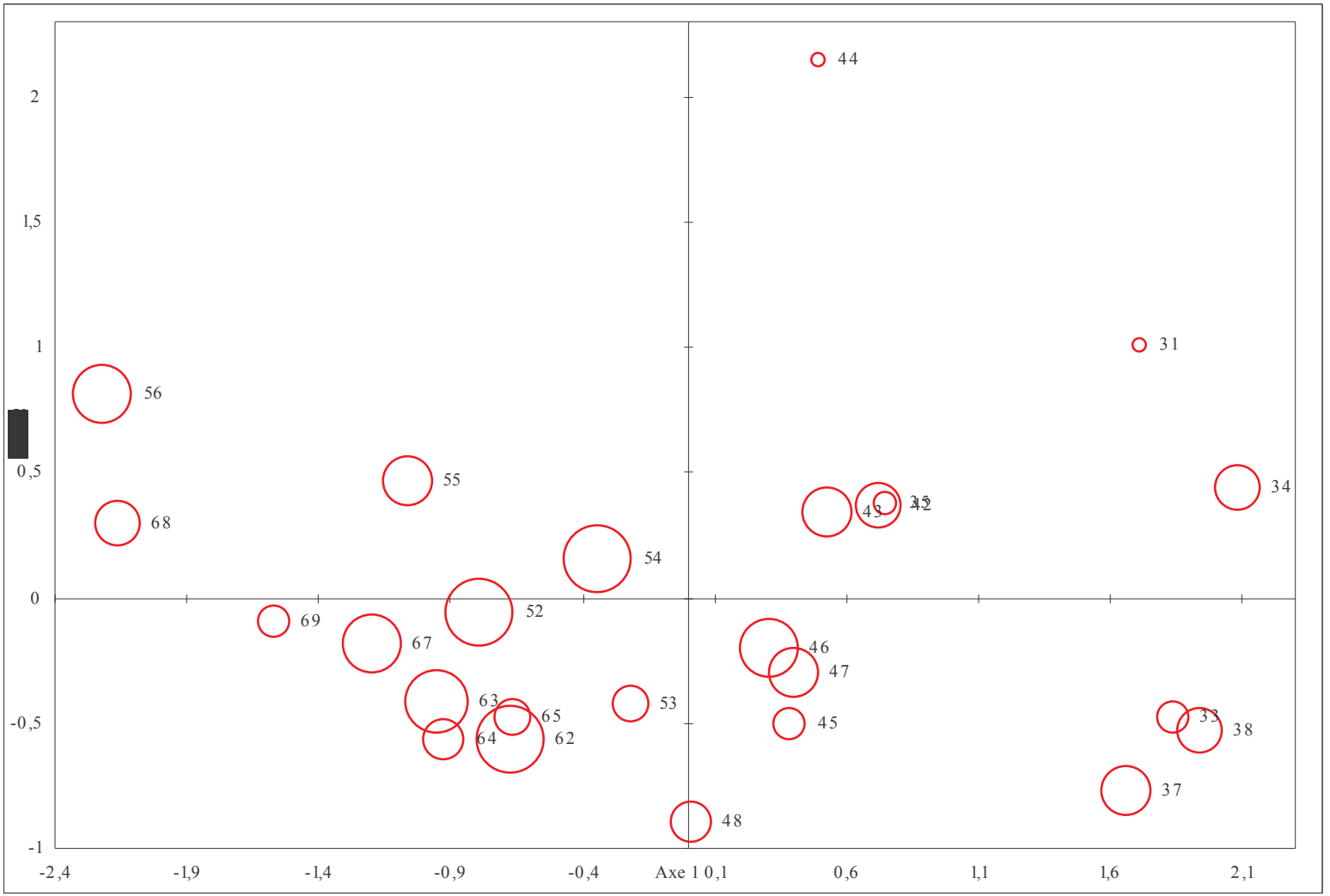
Ex: emploi 2000. salaire(log), Âge et âge de fin d'étude



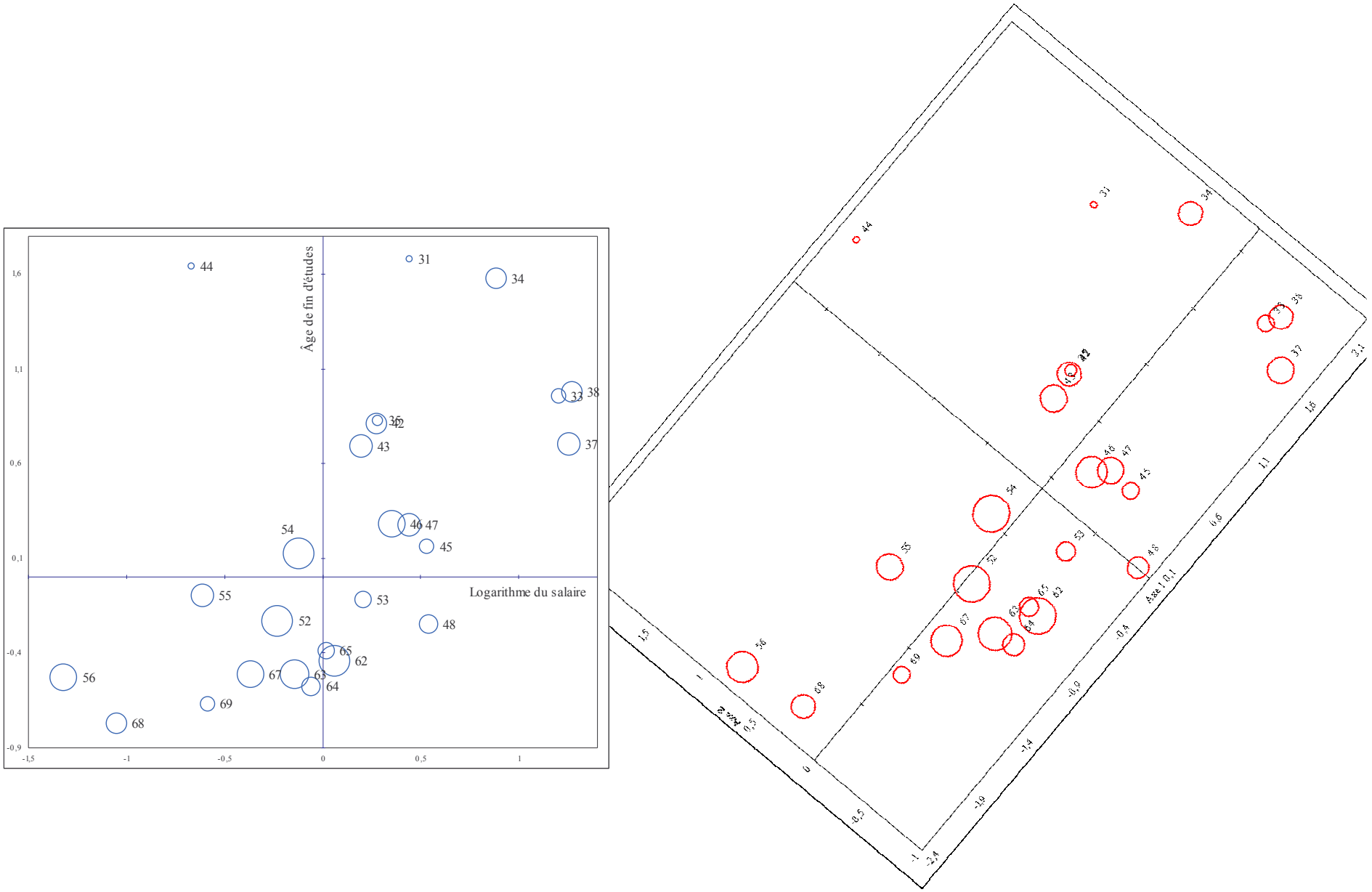
Exemple très simple en deux dimensions



Le plan factoriel issu de l'ACP



C’est le même graphe, mais l’axe horizontal du graphe de droite résume mieux l’info ici.



Entrons un peu dans la méthode

- Inertie : moyenne des carrés des écarts entre les points du nuage et le centre du nuage (point moyen).
- Comment définir une distance ? Distance euclidienne transformée.

  $d^2(a,b)=(x_b-x_a)^2+(y_b-y_a)^2+\ldots(z_b-z_a)^2$

  Arbitraire. D'où le choix préalable de prendre comme « métrique » la variance. On centre et on réduit toutes les variables. la moyenne

  Pour toutes les variables x, on calcule (x-m)/s (la variable moins la moyenne divisée par l'écart-type).

Où ça devient sioux

- Un problème d'optimisation sophistiquée : On cherche à projeter orthogonalement le nuage dans un sous espace de dimension k tel que la distance entre les projections des points du nuage sera maximale.
- Une solution complexe mathématiquement :
  - Cet espace est engendré par les vecteurs propres associés aux k plus grandes valeurs propres de la matrice des coefficients de corrélation des variables.
  - Le premier axe est engendré par le vecteur propre associé à la plus grande valeur propre…
  - Le deuxième axe, etc…
  - Ces axes sont orthogonaux entre eux.

Soyons fous et regardons dans le détail

Données originales

	cs2	lsalred	adfen	agen
	3	9.63	22.51	42.79
X=	4	9.17	20.01	40.07
	5	8.66	18.02	39.41
	6	8.79	16.67	38.33

Données centrées réduites

	cs2	lsalred	adfen	agen
	3	1.31	1.26	1.39
Y=	4	0.24	0.28	-0.04
	5	-0.92	-0.51	-0.39
	6	-0.62	-1.04	-0.96

- Matrice de corrélation

Y'Y=		lsalred	adfen	agen
	lsalred	1.0000	0.9420	0.9199
	adfen	0.9420	1.0000	0.9767
	agen	0.9199	0.9767	1.0000

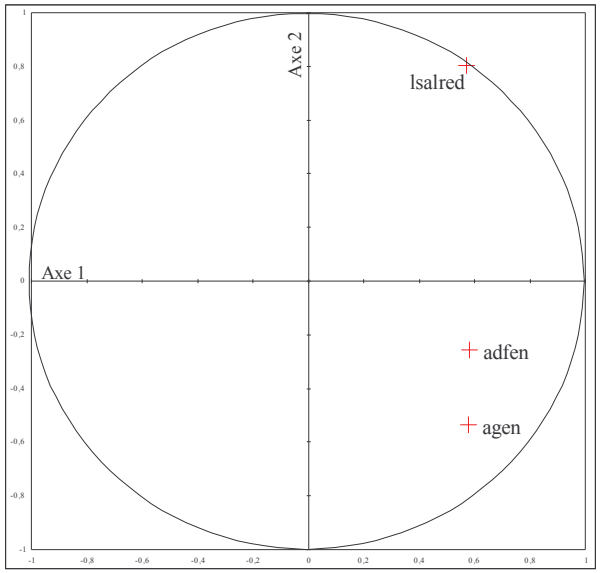
Valeurs propres de la matrice de corrélation

	Valeur propre	Différence	Proportion	Cumulée
1	2.89259588	2.80585359	0.9642	0.9642
2	0.08674229	0.06608046	0.0289	0.9931
3	0.02066183		0.0069	1.0000

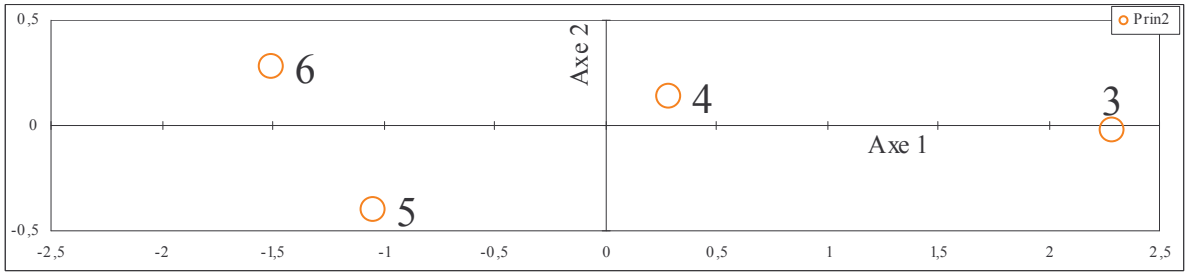
Vecteurs propres

	Prin1	Prin2	Prin3
lsalred	0.571072	0.804163	0.164918
adfen	0.582667	-.255561	-.771484
agen	0.578253	-.536666	0.614503

Espace des variables



Espace des individus



Quelques règles

- On n'a pas le droit dans une ACP de superposer les individus et les variables. Deux sous-espaces de dimensions et de significations différentes.

Variables supplémentaires, classes de variables qualitatives, individus supplémentaires…

- Caractère tautologique de la relation entre variables actives et axes.
- Variables supplémentaires quantitatives : on peut projeter des variables qui n'ont pas participé à la construction des axes dans l'espace des variables. La coordonnée de l'axe est égale au coefficient de corrélation entre l'axe et la variable.
- Variables supplémentaires qualitatives : on représente le point moyen par modalité dans l'espace des individus.
- Individus supplémentaires. On utilise pour les positionner la combinaison linéaire des variables actives définies par l'axe.

L'analyse factorielle de correspondance

- Technique de représentation de tableaux croisés : 2 variables qualitatives croisées entre elles.

Connaît bonus:	Bien	Vague	Non	Total
Cadre Front	9	6	4	19
Cadre Support	6	6	5	17
Technicien	11	14	19	44
Total	26	26	28	80

Une ACP particulière.

On se dote d'une distance la distance dite du chi-deux :

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

AFC (suite)

- Ici on ne centre pas les données, on ne les réduit pas.
- Mais on va chercher les valeurs propres d'une matrice particulière : $M=D_2^{-1}N'D_1^{-1}N$
  - Où N matrice des effectifs du tableau croisé, N' sa transposé, D_1 et D_2 les matrices diagonales contenant sur la diagonale respectivement les totaux en ligne et en colonne.
  - Le terme général de matrice à diagonaliser est le suivant :

$$m_{ij} = \sum_{k=1}^{p} \frac{f_{ki} f_{kj}}{f_{k.} f_{.j}}$$

AFC (suite)

- L'ACP des profils lignes ou des profils colonnes produit des résultats similaires.
- On peut alors superposer l'espace des variables et des individus.

L'ACM

- L'ACM est une AFC particulière…
- Pour l'AFC, on part d'un tableau croisé déjà constitué… pour l'ACM on part du tableau disjonctif complet.
- R=

2	2	4
2	1	3
3	1	2
1	2	4
1	2	1

Z=

0	1	0	0	1	0	0	0	1
0	1	0	1	0	0	0	1	0
0	0	1	1	0	0	1	0	0
1	0	0	0	1	0	0	0	1
1	0	0	0	1	1	0	0	0

- B=Z'Z=

2	0	0	0	2	1	0	0	1
0	2	0	1	1	0	0	1	1
0	0	1	1	0	0	1	0	0
0	1	1	2	0	0	1	1	0
2	1	0	0	3	1	0	0	2
1	0	0	0	1	1	0	0	0
0	0	1	1	0	0	1	0	0
0	1	0	1	0	0	0	1	0
1	1	0	0	2	0	0	0	2

L’interprétation

- Le choix du nombre d’axe
  - En ACP (>1)
  - En AFC et ACM (critère du coude)
- La signification des valeurs propres
  - % d’inertie
  - Mais … pb passage AFC et ACM
- La contribution
- La qualité
- Le cos2
- La valeur test